Rupsa Chakraborty

rupsachakraborty2622@gmail.com | +1 (848)437-1600 | Rupsa-Chakraborty | Rupsa25 |

Summary

– A recent Master's graduate from Rutgers Computer Science, with 8 months of full time experience in **Data Engineering**, I am very passionate about all things data from cleaning to predictive models. Having explored the realm of data through my coursework, I have decided to pursue a career in data engineering and am looking for full time for the same.

Education

Rutgers University,New Brunswick, NJ, USA **(GPA:3.708/4)** Sep 2021 - May 2023

– **Master of Science, Computer Science**

– Coursework:Linear Algebra, Data Structures & Algorithms,Database Management, Data Mining

Vellore Institue of technology, Chennai, India **(GPA:7.85/10)** Aug 2015 - Aug 2019

– **Bachelor of Technology, Computer Science and Engineering**

Technical Skills

– **Programming Languages**: Python,Javascript,Scala,C++, HTML,CSS,Shell Scripting,R

– **Libraries and Frameworks**:Express,NodeJS, NumPy, Pandas, Scikit-learn, TensorFlow,Pytorch,Matplotlib

– **Database and Tools**: SQL Databases,MongoDB, Airflow,Spark,Kafka, Tableau,Github

– **Cloud Services**:AWS Redshift,S3,EC2,Azure Data Lake Storage, Azure Data Factory, Databricks

Experience

Rutgers University,Research Assistant Jun 2022 - Aug 2022

– Implemented and replicated state-of-the-art Active Learning models for object detection on a custom dataset, resulting in a 59.83% reduction in labeling costs for object detection datasets.

– Mapped the dataset to meet model requirements and retrained the models to compare the performance of the novel pipeline against the baseline, achieving a Mean Average Precision (MAP in %) of 26.31 on the first cycle of active learning and 28.6 on the second cycle. **Stack: Python, PyTorch,GNU/Linux Systems**.

Phenom People Private Limited, Data Engineer Dec 2020 - Jul 2021

– Built data pipelines by developing ETL scripts using Python and SQL, scheduling using Apache Airflow, automating data validations and report generations, reducing man hours required for data validations by 20%.

– Ensured adherence pep8 standards and proper coding practices, limiting technical debt to 5%.

– Reduced load time by 10 times by migrating to Amazon Redshift from PostgreSQL, leveraging the Massively Parallel Processing Architecture.**Stack: Python,SQL,Airflow,AWS**.

Academic Projects

Formula 1 Project Jan 2023 - Mar 2023

– Analyzed Formula 1 data, performed data modeling by ingestion, transformation and aggregation of data using **Azure Databricks**,performed data analysis,using **PySpark** and **SQL** queries.

– Used **Azure Data Factory** to schedule the pipeline for automation, incorporated an incremental load architecture for data ingestion,created dashboards for data visualization,taking advantage of IAM roles for permissions for integrating **Azure Data Lake Storage with Databricks**, using the latter for **Data Processing**.

Massive Data Mining Jan 2023 - Mar 2023

– Employed **PySpark** to give friendship on recommendations using data consisting 49995 records.

– Performed **Market basket analysis** to find frequent itemsets on given browsing data consisting 31101 records.

Tableau Projects Dec 2022 - Jan 2023

– Designed Tableau dashboards for netflix titles dataset, goodreads books dataset and Covid -19 data.

What's cooking Kaggle Aug 2018 - Nov 2018

– Developed an SVC-based ML model for cuisine prediction from ingredients, conducted EDA, applied text mining techniques like lemmatization and Tf-idf, achieving an **81.043%** test accuracy with a top leaderboard score of **83.16%**.