# Rupsa Chakraborty

✉ rupsachakraborty2622@gmail.com | in Rupsa-Chakraborty | ⌗ Rupsa25 | Tableau Profile

## Education

**Rutgers University**, New Brunswick, NJ, USA                                    Sep 2021 - May 2023
- **Master of Science, Computer Science**
- Coursework: Linear Algebra, Data Structures & Algorithms,Database for Data Science, Data Mining, Probability & Statistics(Hypothesis Testing),Machine Learning, Database Management Systems

**Vellore Institue of technology**, Chennai, India                                    Aug 2015 - Aug 2019
- **Bachelor of Technology, Computer Science and Engineering**
- Relevant Coursework: Operating Systems, Computer Architecture, NLP, Mathematics

## Technical Skills

- **Programming Languages**: Python,Javascript,Scala,C++, HTML,CSS,UNIX Shell,R
- **Libraries and Frameworks**:Express,NodeJS, NumPy, Pandas, Scikit-learn, TensorFlow,Pytorch,Matplotlib,Scrapy
- **Database and Tools**: SQL Databases, Airflow,Tableau,Github,Apache Spark
- **Cloud Services**:AWS Redshift,S3,EC2,Azure Data Lake Storage, Azure Data Factory, Databricks

## Experience

**Art of Problem Solving,Analytics Engineer**                                    Apr 2024 - Present
- Develop and maintain scalable and robust data pipelines to ensure efficient data extraction, transformation, and loading (ETL) processes,perform data analysis on large scale data using **Python**, **SQL**, and **Dbt**.
- Collaborate with cross-functional teams, including product managers, software developers, and educational specialists, to define data requirements and integrate analytics into user-friendly platforms like **LightDash**.

**World Wide Arc,Co-founder**                                    Dec 2023 - Mar 2024
- Created a website for a job postings website using the **Mongo, Express and Node stack**.

**Diligent Robotics,Clinical Robot Associate**                                    Aug 2023 - Nov 2023
- Troubleshoot and debug a clinical robot in a hospital environment,data collection, automated data extraction efforts using Airflow in combination with **astro cli** in order to streamline deployment, management of Airflow environments.

**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.
- Achieved a MAP score of 28.6% on the second cycle, by mapping the dataset to meet model requirements and retraining the models to compare the performance of the novel pipeline against the baseline. **Stack:PyTorch,Linux**.

**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 by 20% following **CI/CD practices**.
- Reduced load time by 10 times by migrating to **Amazon Redshift** from **PostgreSQL**,optimizing read and write performance through data modeling and design of schemas.
- Lowered **customer onboarding** time by 40% by building an automated job scheduler, in collaboration with cross functional team to retrieve files from **SFTP Server** as soon as created.**Stack: Python,SQL,Airflow,AWS S3,IAM,Git**.

**GVPL Technologies, Data Analyst**                                    Aug 2019 - Oct 2020
- Spearheaded data analysis projects within GVPL Technologies, ensuring a 20% improvement in project delivery timelines through planning and execution.
- Devised data-driven staffing strategies resulting in a 15% reduction in project costs, leveraging offshore technology support to scale department resources according to project requirements. **Stack: Python,SQL,Tableau,AWS,Git**.

## Academic Projects

**Formula 1 Project**                                    Jan 2023 - Mar 2023
- Analyzed Formula 1 data(from Ergast API),performed data ELT (with incremental load architecture) on **Azure Databricks** platform using **PySpark** and **SparkSQL** queries, to analyze performances over a 50 year period,using **Azure Synapse** for Data Warehousing, incorporating **Azure Data Factory** to schedule the pipeline for automation.