

## EDUCATION

### University of Massachusetts Amherst

MS in Computer Science

Sep 2022 – May 2024

### IIT Kharagpur

BTech. + MTech. in Electronics Engineering

Jul 2013 – Apr 2018

## EXPERIENCE

Data Science Intern

Jun 2023 – Aug 2023

### Mathematica

- Conducted causal analysis to explore the relationship between health outcomes and socioeconomic scores across all 3200 US counties. Leveraging statistical techniques and PCA, established a RandomForest model for health score prediction, with an **F1 score of 0.83**, and a Lasso Regression and clustering setup for measuring county similarities for knowledge transfer.
  - Architected an AWS-based framework for the ETL of a **250TB+** dataset from diverse sources (web, S3, MySQL) into Redshift and RDS. Contributed to implementing data quality checks, optimizing performance, and orchestrating with AWS Step Functions.
- Tech Stack:** Python, Pandas, Numpy, Sklearn, Scipy, Transformers, HuggingFace, PySpark, R, Amazon Redshift, AWS Step Functions, AWS Lambdas, Git, AWS CLI, Selenium, Luigi, Missforest, SQL, AWS Glue, Boto, Requests, BS4, URLLIB, Amazon S3

Graduate Machine Learning Researcher

Jan 2023 – Jun 2023

### Adobe Research

- Developed learnable neural image compression algorithms tailored for computer vision pipelines that achieved a 37% reduction in latency. Explored custom multi-task loss functions, fine-tuned model optimization, and precision quantization settings.
- Tech Stack:** Python, Shell, Pytorch, Pytorch Lightning, Weights and Biases, CompressAI, Torchvision, PIL, Matplotlib, CUDA

Software Engineer II - Data

May 2021 – Aug 2022

### Uber

- Developed continuous event streaming analytics pipelines for live and real-time reporting of feedback and sales data leading to 13% reduction in churn rate. Employed Apache Flink for processing data from Kafka streams and loading it into Pinot.
  - Lead engineer responsible for adding new columns, managing data pipelines, writing advanced SQL to create complex and custom data fields, and ensuring data freshness and quality for the **10PB+** central facts table in the Uber Eats domain.
  - Saved \$1.5M** in operational costs by engineering a resource optimization application that identified and optimized expensive Spark and SQL queries using pattern matching and parsing the Spark execution plan, and automatic JIRA ticket assignment.
  - Managed organization's data lake for real-time and batch processing offerings and designed data models for ad-hoc queries.
- Tech Stack:** Python, Java, Scala, SQL, Presto, Spark, Hadoop, Hive, Kafka, Flink, Git, Pinot, Pandas, JIRA, Athena, AWS Data Pipelines, Sagemaker, EMR, RDS, Cassandra, Grafana, YARN, Horovod, HUDI, Tensorflow, Mesos, Docker, MYSQL, Gradle

Senior Data Engineer

Oct 2020 – Apr 2021

### Envestnet Yodlee

- Created a Master Data Management application using Neo4j **graph database** as a single source for time and relationship based data dependencies allowing for a 60% increase in querying speeds and reducing storage by eliminating table joins.
  - Architected data pipelines triggered by creation of materialized views in S3, processed via EMR, and stored in Redshift.
- Tech Stack:** Java, Neo4j, AWS EC2, AWS Data Pipelines, AWS Lambda, S3, Boto, Redshift, Hadoop, Hive, Spark, SQL, CQL

Data Scientist

Jul 2018 – Sep 2020

### SAP

- Built a datalake on Hadoop using Spark and Hive, to analyze application logs (Splunk, ELK) for predictive issue identification.
  - Analyzed seasonality, periodicity, and clustering in application failures to forecast outage costs on Azure and GCP.
  - Deployed code changes and managed fault tolerance and housekeeping across the organizational Hadoop cluster (**150TB+**) using Ambari, Yarn, and Zookeeper. Provisioned data warehouses in Elasticsearch to access the Hive tables in Grafana/Kibana.
- Tech Stack:** Python, SQL, Java, Selenium, Hadoop, Hive, Splunk, Elastic Search, Kibana, Grafana, Spring, Ambari, YARN, Zookeeper, Logstash, Shell, Linux, Maven, Terraform, Ansible, Google Cloud Platform, Microsoft Azure, Pandas, Sklearn

## RESEARCH

### Evaluating inter-agent dynamics of finetuned LLM agents

Sep 2023 – Present

Researched the influence of one LLM agent over another in collaborative and debate scenarios. Constructed agents using Langchain (Llama, GPT, and Vicuna) and provided zero-shot role assignments. Evaluated naturalness using Bert scores and perplexity metrics and in-context learning through dynamic aspect-based sentiment analysis.

**Tech Stack:** Python, Pytorch, Langchain, HuggingFace, GPTQ, AWQ, Accelerate, BERT Score, Transformers, Sklearn, CUDA

## CERTIFICATIONS

**Microsoft Certified: Azure Data Scientist Associate**  
**Neo4j Certified Professional**

Credential ID: H446-0997

Credential ID: 17127043