

EDUCATION

<b>University of Massachusetts Amherst</b> MS in Computer Science	Sep 2022 – May 2024
<b>IIT Kharagpur</b> BTech. + MTech. in Electronics Engineering	Jul 2013 – Apr 2018

EXPERIENCE

<b>Data Science Intern   Mathematica</b>	Jun 2023 – Aug 2023
<ul style="list-style-type: none"><li>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 0.79, and a Lasso Regression and clustering setup for measuring county similarities for knowledge transfer.</li><li>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.</li></ul> <b>Stack:</b> Python, Pandas, Numpy, Sklearn, Scipy, Transformers, HuggingFace, PySpark, Redshift, RDS, SQL, Boto, S3, Git	
<b>Graduate Researcher   Adobe</b>	Jan 2023 – Jun 2023
<ul style="list-style-type: none"><li>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.</li></ul> <b>Stack:</b> Python, Shell, Pytorch, Pytorch Lightning, Weights and Biases, CompressAI, Torchvision, PIL, Matplotlib, CUDA	
<b>Data Engineer II   Uber</b>	May 2021 – Aug 2022
<ul style="list-style-type: none"><li>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 and loading data from Kafka streams into Pinot.</li><li>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.</li><li>Saved \$1.5M in operational costs by engineering a resource optimization application that identified and optimized resource intensive Spark and SQL queries using pattern matching and parsing the AST tree obtained from the Spark execution plan.</li><li>Managed LOB’s data lake (100PB+), on Hadoop and S3, and built data pipelines for real-time and batch processing offerings.</li><li>Wrote data quality, data freshness, and data drift tests for 12% of all deployed data resources, and optimized ETL orchestration.</li></ul> <b>Stack:</b> Python, Java, Scala, SQL, Presto, Spark, Hadoop, Hive, Kafka, Flink, Git, Pandas, Athena, Cassandra, Grafana	
<b>Senior Data Engineer   Investnet</b>	Oct 2020 – Apr 2021
<ul style="list-style-type: none"><li>Created a Master Data Management application using Neo4j NoSQL 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.</li><li>Architected data pipelines triggered by creation of materialized views in S3, processed via EMR, and stored in Redshift.</li></ul> <b>Stack:</b> Java, Neo4j, AWS EC2, AWS Data Pipelines, AWS Lambda, S3, Boto, Redshift, Hadoop, Hive, Spark, SQL, CQL	
<b>Data Engineer   SAP</b>	Jul 2018 – Sep 2020
<ul style="list-style-type: none"><li>Built a datalake on Hadoop using Spark and Hive, to analyze application logs (Splunk, ELK) for predictive issue identification.</li><li>Analyzed seasonality, periodicity, and clustering in application failures to forecast outage costs on Azure and GCP.</li><li>Deployed code changes and managed fault tolerance and housekeeping across the organizational Hadoop cluster (150TB+).</li><li>Created reports and visualizations in Splunk, Kibana, and Grafana for monitoring application metrics and system health.</li></ul> <b>Stack:</b> Python, SQL, Java, Selenium, Hadoop, Hive, Splunk, Elasticsearch, Grafana, Ambari, Zookeeper, Shell, Terraform, Azure	

SELECTED RESEARCH

<b>Evaluating inter-agent dynamics of finetuned LLM agents</b>	Sep 2023 – Present
Studied and quantified influences in multi-agent debate setup. Employed Langchain to chain responses (Llama, GPT, and Mistral) and assigned roles through fine-tuning and zero-shot instruction prompting. Experimented with Chain-of-Thought to direct the conversation, and evaluated synthetic data generated on perplexity, drift, and coherence. Quantized the models for inference.	
<b>Tech Stack:</b> Python, Pytorch, Langchain, HuggingFace, GPTQ, AWQ, Accelerate, BERT Score, Transformers, Sklearn, CUDA	

CERTIFICATIONS

<b>Microsoft Certified: Azure Data Scientist Associate</b>	Credential ID: H446-0997
<b>Neo4j Certified Professional</b>	Credential ID: 17127043
<b>Scala and Functional Programming Essentials</b>	Udemy Certification
<b>Tensorflow 2.0 : Deep Learning and Artificial Intelligence</b>	Udemy Certification

HONORS

- Selected as a beta tester for Manning publication of "Real-time Stream Processing with Kafka and Spark"
- Graduate Teaching Assistant for Advanced Machine Learning (Fall 2023) and Computer Vision (Spring 2024) at UMass Amherst
- Taught Data Science and Data Engineering industry skills to experienced professionals as an instructor at Scaler Academy