KARTIK GUPTA

kartikgt23@gmail.com
github.com/kartikgt
linkedin.com/in/kartikgupta23
(413) 468-0353

EDUCATION

University of Massachusetts Amherst *Sep 2022 – May 2024*
MS in Computer Science *GPA - 3.8*

IIT Kharagpur *Jul 2013 – Apr 2018*
BTech. + MTech. in Electronics Engineering

SKILLS

Certificates	Microsoft Certified Azure Data Scientist Associate, Neo4j Certified Professional
Languages and Tools	Python, Java, Scala, R, Golang, Gradle, SQL, Cypher, Linux, Git, Shell, Kubernetes, Boto, YARN
Data Science	Pandas, Numpy, Scikit, Statistics (Bayesian and Stochastic), Scipy, Regression, Clustering, SVD
Data Engineering	Spark, PySpark, Flink, Kafka, Hadoop, Hive, MapReduce, Databricks, Snowflake, ElasticSearch, DBT
AI and ML	Pytorch, HuggingFace, Langchain, Lightning, Tensorflow, OpenCV, LLM, FSDP, Accelerate, CUDA

EXPERIENCE

Mathematica *Jun 2023 – Aug 2023*
Data Science Intern *Chicago, USA*

- 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 feature-space county distances. *(Scikit-learn, Pandas)*
- 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.
- Designed data warehouses within Redshift to house critical Medicaid and private insurance rates for both individual and bundled medical procedures. Performed data analysis using PySpark, as well as deployment of data pipelines via AWS Lambda.

Adobe Research *Jan 2023 – Jun 2023*
Graduate Machine Learning Researcher *Amherst, USA*

- 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.
- Employed PyTorch Lightning framework for modular code and streamlined training pipelines in multi-GPU setup. *(CUDA, FSDP)*

Uber *May 2021 – Aug 2022*
Software Data Engineer II *Remote*

- Developed continuous event streaming analytics pipelines for live and real-time reporting of feedback and sales data. Employed Apache Flink for processing data from Kafka streams and loading it into key-value data stores. *(Java, Scala, Flink)*
- 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. *(Java, Scala, Python, SQL, Spark)*
- Reduced HDFS **disk space usage by >60%** by converting AVRO and ORC tables into Parquet format and replacing GZIP and SNAPPY compression with ZSTD. Performed periodic housekeeping for managing the diskspace and namespace usage.

Envestnet Yodlee *Oct 2020 – Apr 2021*
Senior Data Scientist *Bangalore, India*

- 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.
- Wrote botocore scripts to automate AWS Data Pipelines to transform new data files from S3 to EMR using AWS Lambda.

SAP *Jul 2018 – Sep 2020*
Data Scientist *Bangalore, India*

- Built a SSOT datalake on HDFS 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.

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. *(Langchain, HuggingFace, Pytorch)*

Advanced Stream Processing using Kafka and Spark *Aug 2023 – Oct 2023*
Selected as a beta tester for a new Manning liveProject dealing with Kafka and Spark stream processing. Specifically evaluated sections on advanced stream processing, and provided feedback on applicability to standard use cases. *(Java, Spark, Kafka)*