Roshni Pal

Boston, MA • roshnipal28@outlook.com • (774) 636-8852 • LinkedIn • GitHub

EDUCATION

University of Massachusetts Dartmouth — May 2025
Master of Science, Data Science — GPA: 3.9/4.0
Acropolis Technical Campus — July 2021
Bachelor of Engineering, Computer Science and Engineering — CGPA: 8.48/10.0

PROFESSIONAL EXPERIENCE

University of Massachusetts Dartmouth
Research & Development Assistant — **Sept 2024 - Present**

- Developed a scalable machine learning pipeline to predict turbulent fluid dynamics, processing 1.8 billion velocity records from 12 experimental datasets.
- Designed a variational autoencoder (VAE) for dimensionality reduction into a 47-dimensional latent space and implemented a transformer model for time-series predictions.
- Leveraged SLURM to parallelize workflows across 1000+ jobs, utilizing transfer learning, data preprocessing, and denormalization.
- Achieved high predictive accuracy, reducing computational overhead and enabling advanced analysis of turbulence dynamics.

Rakuten Symphony
Data Engineer — **July 2021 - July 2023**
Project 1: Rakuten Coverage Platform (RCP)

- Migrated product to microservices architecture using RESTful APIs, boosting scalability, configurability, and fault tolerance.
- Designed and optimized ETL pipelines with Apache NiFi and Airflow, ensuring seamless data transformation, robust data management, and compliance with governance standards.
- Achieved a 74% improvement in MySQL and Cassandra query performance through indexing and partitioning techniques.
- Developed Kafka-based streaming jobs to process real-time network data, integrating geospatial data layers with Google Maps APIs, improving coverage prediction accuracy by 25%.

Project 2: Sympulse—Network Monitoring Tool

- Engineered batch and real-time streaming pipelines using Apache Spark and Kafka, handling 5,000+ events/sec and managing 5 TB of data on the cluster.
- Automated extraction of 300+ KPIs and deployed MLOps pipelines for network performance predictions, enhancing proactive issue resolution by 80%.
- Utilized AWS services (EC2, S3, RDS, Lambda) to manage infrastructure and implement serverless functions for data transformation and validation.
- Consolidated Spark jobs into Kubernetes pods, improving cluster efficiency by 60%, reducing memory usage by 30%, and automating workflows with shell scripting to cut manual overhead by 90%.

Rudra-x Software Solutions
Data Analyst — **Jan 2020 - June 2020**

- Analyzed large volumes of claims, premium, and policyholder data to identify actionable insights, leading to a 25% improvement in data-driven decision-making for policy pricing and claims optimization.
- Utilized Python and SQL to analyze trends in claims and automate data processing tasks, improving decision-making efficiency by 30% and reducing processing time by 20%.
- Collaborated with business intelligence teams to validate 10+ BI dashboards and reports, ensuring precise representation of metrics and compliance with healthcare regulations.

PROJECTS

Master's Thesis: Unified Framework for Motion Understanding: From 3D Skeletons to Multi-Modal Reasoning

- Existing systems lacked integration and precision for tasks like 3D pose estimation, action recognition, and motion captioning across diverse data types, including videos and skeleton sequences.
- Develop a unified framework combining motion analysis and multi-modal reasoning to achieve high performance on key benchmarks.
- Integrated MotionBERT's DSTformer and MotionLLM's Vision-Language Translator, leveraging 2D-to-3D pretraining and fine-tuning on motion-video-text datasets using AMD Radeon W7900 GPU.
- Achieved 37.5 mm MPJPE on Human3.6M and a 45% improvement on MoVid-Bench, advancing accuracy in pose reconstruction, action recognition, and motion captioning.

Automated Data Extraction and Classification with Machine Learning and OpenAI API

- Designed a solution to streamline data ingestion and improve classification efficiency across diverse file formats (xlsx, pdf, docx, txt).
- Developed an API integrating ETL pipelines and OpenAI API for automated data extraction, tagging, and dynamic prompts, and built machine learning models for report classification.
- Realized a 40% boost in processing efficiency and attained 91% classification accuracy, significantly improving reporting workflows.

SKILLS

- **Language/Libraries:** Python, Java, R, NumPy, Pandas, Matplotlib, TensorFlow, XGBoost, scikit-learn
- **Database:** MySQL, MSSQL, Cassandra, HBase, Apache CouchDB
- **Distributed System:** Apache Spark, Spark Streaming, Kafka, Docker, Kubernetes, Hadoop, AWS (QuickSight, Glue, MinIO, S3, Lambda, DynamoDB), Azure (Data Lake Gen 2, Data Factory, Databricks), Hadoop, CI/CD, Prompt Engineering
- **Machine Learning/AI Models:** Linear & Logistic Regression, Random Forests, Decision Trees, Neural Networks, Deep Learning, NLP
- **Data Visualization/Methodologies:** Power BI, Tableau, Advanced Excel (Pivot Tables, Lookup), Git, Agile (Scrum, JIRA)