Kirthivasan Pandurangan Neelavathi

812-764-0888 | kirpandu@iu.edu | kirthivasan-pn | KirthivasanPN-hash.github.io | (Open to relocation)

EDUCATION

Indiana University — Bloomington, Indiana
Master of Science in Data Science — August 2023 – May 2025

St. Joseph's College of Engineering — Chennai, India
Bachelor of Engineering in Computer Science – GPA: 3.8/10 — August 2019 – April 2023

TECHNICAL SKILLS

Academic Courseworks: Statistics, Applied Algorithms, Database Concepts, Machine Learning in Econ, Data Mining, Big Data
Programming Languages: Python, R, SQL, Cypher, JavaScript
Python Libraries: Pandas, NumPy, Matplotlib, Seaborn, XGBoost, OpenCV, TensorFlow, HuggingFace
ML Algorithms: Linear Regression, Random Forest, Decision Tree, k-means clustering, SVM, Naive Bayes, Neural Networks
Data Mining Techniques: TF-IDF, Word2Vec, feature extraction, clustering algorithms
Database & Big Data: MySQL, PostgreSQL, MongoDB, Apache Spark, AirFlow, Neo4j, Databricks, Docker, AWS

WORK EXPERIENCE

Graduate Research Assistantship — August 2024 – present
Indiana University — Indianapolis, Indiana

- Programmed a data collection pipeline using BeautifulSoup and Selenium to automate the extraction of 1,000+ data points from online sources, improving data acquisition efficiency and reducing manual data collection effort.
- Built a text processing pipeline using PyMuPDF and spaCy Sentencizer to extract and analyze sentence-level representations, reducing time complexity for extraction process and enabling faster analysis for datasets.
- Incorporated LangChain for chunking, transforming multi-sentence contexts into chunks by retaining critical information for further analysis. Reduced data complexity and improvised subsequent analytical tasks.
- Streamlined data transformation processes using tqdm, generating embeddings for over 500 sentences and implementing retrieval techniques to optimize data processing pipelines for faster insights.
- Integrated Large Language Models for PDF summarization and key insight extraction, content analysis, keyword identification, and contextual understanding, reducing manual effort and enhancing data processing efficiency.

Software Engineer Internship — June 2024 – August 2024
Food Fight — Remote

- Created RESTful APIs using FastAPI, enabling seamless application-database communication and achieving a 15% increase in API responsiveness. Adhered Pydantic models for data validation, database operations, and reducing error rates by 10%.
- Enabled SQLAlchemy for Pythonic database interactions, enabling seamless execution of queries using an ORM. Optimized query performance, reducing execution time and improving database efficiency, leading to a 10% decrease in error rates.
- Engineered an user matching recommendation system by leveraging cosine and Jaccard similarity metrics. Analyzed multidimensional user data including preferences, behavioral patterns, and historical interactions for personalized recommendations.
- Leveraged AWS Lambda to support serverless tasks, optimizing resource usage and reducing idle server costs by 20%. Integrated AWS CloudWatch for real-time log monitoring and alerting, assisting in issue detection and resolution.

PROJECTS

Automated Movie Data Transformation Pipeline

- Designed and coded a robust ELT pipeline using Docker containers, automating the extraction, transformation, and loading of large-scale movie dataset. Performed Orchestration via Airflow for optimizing task scheduling and error handling.
- Integrated DBT for data transformation, utilizing Jinja templates and macros to create SQL models for filtering the movie dataset. Executed transformations, complex aggregations, joins, and data enrichment techniques, to ensure optimization.
- Configured a cron job for automated pipeline execution, streamline the entire workflow, minimizing manual intervention, reducing errors, and operational efficiency by maintaining regular updates and data processing intervals.

Market Pulse: Real-Time Stock Data Pipeline

- Initiated AWS EC2 to manage Kafka brokers and producer/consumer communications, enabling real-time data streaming services. Configured Zookeeper for efficient cluster coordination and fault-tolerant operations.
- Implemented Boto3 SDK to interact with AWS services via Python, facilitating real-time data ingestion from Kafka into S3 buckets while ensuring swift preparation for cataloging in RDS via AWS Glue.
- Set up a real-time stock price pipeline and AWS Glue crawlers to load and structure data in Athena, enabling real-time stock market analysis through SQL queries and insights generation.

Aviation Analytics Pipeline on GCP

- Led the development of an ETL pipeline using Google Cloud Platform, streamlining data processes and increasing efficiency by 25%. Utilized Cloud Storage for data warehousing, and workflow orchestration, resulting in a scalable data infrastructure.
- Architected a secure Virtual Private Cloud (VPC) for cluster deployment, optimizing large-scale aviation data processing, Cloud IAM for granular access control, and data handling efficiency by 30%.