DEEPIKA ALAGIRISWAMY PANNEERSELVAM

Chicago, IL | (630) 873-0087 | deepikapanneerselvam6@gmail.com | https://www.linkedin.com/in/deepika-panneerselvam/ |
https://github.com/DeepikaPanneer | https://deepikapanneer.github.io/Portfolio/

EDUCATION

Illinois Institute of Technology, Chicago, IL May 2025

Master of Data Science, GPA: 3.90

- Coursework: Data Preparation and Analysis, Big Data Technologies, Machine Learning, Regression, Statistics, Advance Database Organization.

Coimbatore Institute of Technology, Coimbatore, India May 2021

Bachelor of Technology, Information Technology, GPA: 3.80

- Coursework: Data Structures, Database Management System, Object Oriented Programming, Cloud Computing, Data Mining and Warehousing.

EXPERIENCE

Urban Institute - Washington DC, MD

Data Engineering Intern June 2024 - November 2024

- **Modernized an 8-year-old data analytics system** by automating version updates for RStudio, R, and EMR through custom API calls. Seamlessly integrated updates into the legacy infrastructure and delivered the full upgrade two months ahead of schedule. This reduced manual maintenance time by 50% and improved environment stability, allowing analysts to run workloads 25% faster.
- **Designed and implemented AWS Step Functions** to orchestrate EMR cluster creation and EC2 instance provisioning, streamlining automation and reducing setup time from hours to minutes. This enabled the team to accelerate troubleshooting and reduced the manual intervention.
- Built a CI/CD pipeline using **AWS SAM** and **GitHub Actions**, fully automating deployments for entire project, cutting the deployment time by 60%

Dun & Bradstreet - Hyderabad, India

Big Data Engineer August 2021 - July 2023

- Improved **80% of datasets** by adding key business features and restructuring data models, resulting in smoother access for analytics and reporting.
- Deployed **3 end-to-end ETL pipelines** into production using **PySpark, Snowflake, EMR, S3** and **Airflow**, improving data delivery with minimal support.
- Orchestrated workflows using **Apache Airflow** and contributed to **Snowflake data modeling** to streamline pipelines and improve query efficiency.
- **Reduced production failure by 30%** by identifying and fixing high-priority bugs through **debugging**, **performance tuning** and **optimization**.
- Handled diverse file formats—including nested **JSON**, **CSV**, and **Parquet**—within automated **data ingestion** workflows using **AWS Lambda.**
- Set up end-to-end **CI/CD pipelines** using **Jenkins**, ensuring version control and automated testing, to enable, faster deployments and reduce errors.
- Created **SQL scripts** to validate data pipelines, ensuring **accuracy**, **compliance**, and **data governance**- reducing manual checks by approximately 40%.

Data Engineer Intern February 2021 - August 2021

- Documented **6 ETL workflows** built with **PySpark** and **Snowflake**, improving knowledge transfer and reducing onboarding time for new engineers.
- Monitored production jobs using **AWS CloudWatch** and SQL validation, reducing data issues and potential downtime by approximately 50%.
- Collaborated with **cross-functional stakeholders** and senior engineers to develop **scalable data pipelines** for high-impact and critical tasks.

SKILLS

Programming Languages & Libraries: Python, PySpark, Pandas, NumPy, R, Shell Scripting
Big Data Technologies & Databases: Apache Spark, Databricks, Hadoop, SQL, Snowflake, MySQL
Cloud Services: AWS: Lambda, Step Functions, SAM, EMR, S3, CloudFormation; **GCP:** BigQuery, Dataproc, Cloud Storage, Cloud Functions
Automation & Orchestration: Apache Airflow, Jenkins, Git, GitHub, GitHub Actions, Jira, Docker
Data Analysis & Visualization: Microsoft Excel, Tableau

PROJECTS

Stock Market Data Pipeline using Kafka and AWS March 2025

- A real-time data processing pipeline that simulates live trading data using Python, streams it into an Apache Kafka topic hosted on an AWS EC2 instance, and securely stores the data in Amazon S3 buckets for persistent storage and further analysis.
- The stored data in S3 is crawled and cataloged using AWS Glue, enabling efficient structured querying and analysis using Amazon Athena.
- **Tech Stack:** Python, Pandas, Apache Kafka (on AWS EC2), AWS S3, AWS Glue (Crawler + Data Catalog), AWS Athena.

Multilabel Predictions on Academic Articles, Illinois Tech, Chicago, IL April 2024

- Multi-label classification model that processes academic articles using TF-IDF vectorization and BERT to train the models for predicting article's subjects.
- Model performance is evaluated with precision, recall, and F1-score metrics; Matplotlib visualizations to understand label distribution and EDA.
- **Tech Stack:** Machine Learning, NLP, Python, Pandas, NumPy, BERT, scikit-learn, Matplotlib.

Predictions for the 2023 ICC Cricket World Cup, Illinois Tech, Chicago, IL November 2023

- A machine learning model in R to predict match outcomes for the 2023 ICC Cricket World Cup using historical data from 10 international teams.
- Trained Random Forest models incorporating team statistics and factors like venue, toss outcome, and pitch conditions to enhance prediction accuracy.
- Performed exploratory data analysis with Tidyverse and visualized performance trends, correlations, and feature importance using ggplot.
- **Tech Stack:** R, Random Forest, ggplot, Tidyverse

CERTIFICATIONS

- **AWS Certified Data Engineer – Associate** *(August 2024)*
- **Google Cloud Certified Cloud Digital Leader** *(February 2025)*