Deepika Alagiriswamy Panneerselvam

Chicago, IL | (630) 873-0087 | deepikapanneerselvam6@gmail.com | LinkedIn| Github | Portfolio

EXPERIENCE

Urban Institute - Washington DC, USA

Data Engineering Intern June 2024 - November 2024

- **Modernized the 8-year-old EMR cluster provisioning system** by automating version updates for RStudio, R, and EMR via custom APIs, reducing manual effort by 50% and improving workload speed by 25%. Delivered the upgrade two months early, enhancing system stability and efficiency.
- **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, SQL, PySpark, Pandas, NumPy, R, Shell Scripting.
Big Data Technologies & Databases: Apache Spark, Databricks, Hadoop, Snowflake, MySQL.
Cloud Services: AWS - Lambda, EMR, Step Functions, SAM, S3, CloudFormation.
Automation, Orchestration & Visualization: Apache Airflow, Jenkins, Git, GitHub, GitHub Actions, Jira, Docker, Microsoft Excel, Tableau.

PROJECTS

Stock Market Data Pipeline using Kafka and AWS March 2025

- Designed a real-time pipeline in Python to simulate live trading data, stream via Kafka, and store securely in S3 buckets.
- Used AWS Glue to crawl and catalog S3 data, enabling fast, structured querying and analysis using Amazon Athena with ease.
- **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

- Built an R-based machine learning model to predict ICC World Cup 2023 outcomes using historical team data and visualized insights.
- Trained Random Forest models including team stats, venue, toss, pitch data to improve prediction accuracy and performance of match outcomes.
- **Tech Stack:** R, Random Forest, ggplot, Tidyverse.

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.

CERTIFICATIONS

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