Tarun Kumar

☎ 8167166112 ✉ tarunkumarw2@gmail.com in LinkedIn GitHub

Experience

USAA — Jan 2024 - Present
Senior Data Engineer — San Antonio, TX, USA

- Designed and automated **scalable ML workflows** for deployment and retraining using **Apache Airflow DAGs, GitLab CI/CD, and Docker**, ensuring high reliability and reducing model retraining time by 30%.
- Implemented an end-to-end data integration, transformation, and analytics framework using **Microsoft Azure Data Factory and Microsoft Fabric** to streamline the reporting process, achieving a 37% reduction in report generation time, enhancing data accuracy by 25%, and an 8% increase in revenue.
- Developed and **deployed machine learning models** for fraud detection across various banking stages, including member authorization and card/non-card transactions, resulting in a 15% reduction in fraudulent activities.
- Performed **reporting and monitoring tasks** for model pipelines in the fraud detection department, ensuring operational efficiency and achieving 9% enhancement with accuracy in fraud detection alerts.
- Developed **Automated data pipelines** for refining unstructured and semi-structured data, implementing custom data transformations that reduced processing time by 50% and improved data quality.
- Implemented end-to-end **Generative AI** workflow leveraging **LLMs** via **Azure OpenAI Service**, utilizing **LlamaIndex** and **LangChain** for **document processing**, and automating the **end-to-end workflow** with **Apache Airflow**. Implemented **AI agents** to extract relevant information based on user queries and **deployed** the solution on **Azure Cloud**.
- Developed and deployed machine learning models within Model Development Control and Model Risk Management frameworks, and conducted rigorous model validation, reducing validation time by 20%.
- Designed and implemented risk-aware fraud detection models using supervised and unsupervised learning, integrating bias detection techniques and fairness audits to ensure regulatory compliance, reducing false positives by 15%.
- Built and maintained a reusable feature engineering library in **Python**, leveraging **AWS SageMaker Feature Store**, optimizing feature selection and transformation pipelines with **Pandas, PySpark, and Scikit-learn**, improving model accuracy by 18% and reducing feature computation time by 30%.
- Designed and developed **Python and Spark** scripts for data modeling, integrating data from multiple sources to analyze and transform customer conversation patterns, leading to a 20% improvement in customer insights.
- Successfully migrated databases, repositories, and **delta tables from Hadoop to Amazon S3**, ensuring a seamless data transition and improving data accessibility by 40%.
- Developed and integrated unsupervised algorithms into the production system, enabling the identification of optimal cluster numbers for trading data, which facilitated category merging and reduced dimensionality by 30%.
- Utilized **Azure Synapse Analytics** to integrate and analyze large datasets, enabling real-time insights and improving decision-making processes.
- Implemented data storage solutions using **Azure Cosmos DB**, optimizing data retrieval and ensuring low-latency access for real-time applications.

Brown Brothers Harriman — February 2023 - January 2024
Senior Data Engineer — New York, NY, USA

- Designed and deployed scalable, fault-tolerant systems on **Azure**, building **end-to-end data pipelines** with **Data Factory and Databricks** for seamless data ingestion from on-premises to **Azure SQL Database**, while performing incremental and full data loads.
- Configured **Spark streaming** to receive real-time data from **Kafka** to store it in **HDFS**, and Implemented Generative AI applications utilizing **Large Language Models (LLMs)** and **Retrieval-Augmented Generation (RAG) systems**, leveraging **LlamaIndex and LangChain** for efficient data retrieval and integration with **Pinecone (vector database)** that improved customer engagement metrics by 25% and reduced response times by 30%.
- Engineered systems using **Python**, **SQL**, and machine learning algorithms like **Scikit-learn** for anomaly detection, reducing issues by 45% and boosting accuracy by 25%.
- Engineered executive-level **AWS QuickSight** visualizations and **Tableau** reports using optimized **SQL** queries, simplifying complex data trends, reducing dashboard loading time by 35%, and improving quarterly performance tracking by 30%.
- Designed a workflow utilizing **Apache Airflow** to trigger and load 117 models in parallel, optimizing runtime by 71% and increasing efficiency by 50%.

Millipore Sigma, Merck Group — January 2022 - January 2023
Senior Data Engineer — St. Louis, MO, USA

- Engineered intent recognition deep learning models that resulted in an annual cost savings of $350,000 by enhancing chatbot intelligence. Integrated advanced ML/AI algorithms that effectively reduced customer churn by 13% and increased click-through rates by 18%, employing the **ALBERT model** for user intent classification, achieved 91% on production data.
- Developed and optimized **ETL** processes to efficiently extract, transform, and load data from diverse sources into a centralized data warehouse, ensuring high data quality and accessibility for analytics. Utilized **AWS Glue and Lambda**

to automate data workflows and enhance data pipeline performance.

- Designed and orchestrated data pipelines using **Python, Apache Airflow, and SQL**, achieving a 100% automation rate with **Docker** to reduce manual intervention and errors, while architecting advanced AWS workflows for **Redshift** loading and **QuickSight** dashboards to automate transaction extraction and enhanced reporting efficiency by 50%.
- Executed thorough data cleansing, duplicate handling, and transformations using Azure Data Factory and Databricks. Conducted in-depth root cause analysis with Python and Pandas to reconcile mismatched records across millions of entries, resulting in a 40% improvement in accuracy and ensuring data integrity for critical business operations.

Dun and Bradstreet — June 2020 - December 2020
Senior Data Scientist — Chennai, India

- Collaborated with business stakeholders to gather data requirements and developed **Spark scripts** using **Python APIs** to import and process raw files from **S3** into **Spark Dynamic Frames**, converting them to DataFrames for transformations and implementing data quality metrics scripts to monitor daily partition files.
- Created **PySpark scripts** to merge static and dynamic files, cleanse the data, and convert **Hive/SQL queries** into **Spark transformations using Spark DataFrames in Python**, ensuring data integrity and quality for target tables.
- Architected advanced AWS workflows, including **Redshift** loading and **QuickSight** dashboards, to automate transaction extraction, increasing reporting efficiency by 50% and significantly accelerating business intelligence tasks and improving operational efficiency.
- Enhanced the 'Analytics Studio' platform by building advanced risk assessment models, including propensity, probability of default (PD), loss given default (LGD), and financial stress score (FSS) using DUNS data, which contributed to a 32% increase in product revenue for the fiscal year.

INSOFE — March 2018 - May 2020
Data Scientist — Bangalore, India

- Developed and deployed an **ABSA model for sentiment analysis** of automobile reviews and implemented semantic segmentation using the **Faster-RCNN algorithm** for tumor detection, achieving high accuracy in real-time applications.
- Utilized **advanced pre-trained convolutional neural networks** for diabetes severity detection, and customized models to enhance F1-score by 8% and achieved 93% accuracy for real-time data.
- Conducted requirements gathering, performed daily data extraction from multiple tables, and transformed raw data using **Azure Data Factory and Databricks** with **Python, Pyspark, and SPARK SQL**, while **optimizing SQL queries in MySQL** to reduce execution time by 20%.
- Managed data engineering workflows by scheduling and monitoring pipelines, leveraging **Apache Airflow** to improve **ETL** processes, migrating databases to **AWS** for increased scalability and cost reduction, and implementing automated backup solutions to enhance reliability and reduce downtime by 17%.

Larson and Toubro Technology Services — April 2017 - February 2018
Associate Software Engineer — Hyderabad, India

- Enhanced data transfer efficiency by loading and processing data from **UNIX** file systems to **HDFS** using **Hive UDFs and Sqoop**, and improved query performance by transferring data between **DB2 and HBase**.
- Developed resume parser to analyze candidate profiles for Internal Job Postings, achieving a 98% screening precision through real-time relevance ranking, and migrated the on-premises **HDFS ecosystem** to **AWS Cloud**, transitioning **data lakes and warehouses** to **AWS S3 and Redshift**, while utilizing **Apache Spark** for large dataset analysis.
- Designed automated data pipelines with **Python**, **Apache Airflow and SQL**, achieving 100% automation with **Docker**, and Architected **AWS** workflows that improved reporting efficiency by 50%.

Skills

- **Hadoop Ecosystem:** Hadoop, Distributed Systems, PySpark, MapReduce, Hive, YARN, Kafka, Zookeeper, Airflow
- **Databases/Querying:** Oracle, My SQL, SQL Server, PostgreSQL, HBase, Snowflake, Cassandra, MongoDB, T-SQL
- **Cloud Computing:** Azure Data Factory, Azure Databricks, Azure Synapse Analytics, Azure Data Lake storage, Azure Cosmos DB, Microsoft Fabric, Amazon Web Services (AWS), Amazon S3, AWS Glue, AWS Lambda, Amazon Redshift
- **BI tools:** Tableau 9.1, Power BI
- **Scripting Languages:** Unix, Python, Linux
- **Generative AI/Machine Learning:** LLM, RAG, RHLF, Llama Index, LangChain, GPT's, Embeddings, Transformers, BERT, Deep Neural Networks, Supervised and Unsupervised Learning, Linear Regression, Logistic Regression, Time Series, Clustering, Naive Bayes, Decision Trees, Support Vector Machines, XGBoost, Hyper Parameter Optimization, Data preprocessing, Spacy, Sentiment Analysis, Text Classification
- **SDLC Methodology:** Agile, Scrum, Waterfall

Certifications

- AWS Solutions Architect - Professional (Link)
- DP-600: Microsoft Fabric Analytics Engineer Associate (Link)

- DP-203: Azure Data Engineer Associate (Link)
- PGP in Data Science and Big Data Analytics in INSOFE (Certified by LTI of Carnegie Mellon University) (Link)
- Udemy – Apache Flink and Kafka End-to-End Streaming Project (Link)
- Udemy – Azure Databricks (Link)

Education

Master of Science in Computer Science
University of Central Missouri, Warrensburg, MO, USA

Bachelor of Technology in Electrical and Electronics Engineering
KL University, Guntur, India