Venkata Sai Tarun S

Houston, TX | +1 (713) 659- 9389 |saitarun.s42@gmail.com | LinkedIn | GitHub

SUMMARY

AWS Certified **Data Engineer** with **around 5 years** of experience delivering scalable AI/ML data architectures, high-performance ETL/ELT pipelines, and production-grade machine learning workflows across consulting, financial services, and enterprise environments. Expert in designing and optimizing cloud-native data platforms on AWS, Azure, and GCP, leveraging Databricks, Spark, Kafka, and distributed computing to support real-time analytics and large-scale data processing. Proven success in operationalizing ML models with MLOps, MLflow, Docker, and Kubernetes, improving model accuracy, reducing latency, and accelerating deployment cycles. Adept at building robust data lakes, data warehouses, and streaming pipelines, and delivering actionable insights through BI solutions in Power BI, Tableau, and Python. Strong domain expertise in NLP, Generative AI, statistical modeling, and advanced analytics, with a consistent track record of driving data-driven decision-making, enhancing system reliability, and optimizing compute and cloud costs in fast-paced, mission-critical environments.

SKILLS

Programming Languages: Python, SQL, PySpark, R, Java
Machine Learning & AI: Scikit-learn, PyTorch, TensorFlow, XGBoost, Keras, Dask, NLP, Generative AI, Risk Stratification, Responsible AI
ML Operations (MLOps): Model Deployment, Model Serving, Data Validation, Monitoring, MLflow
Cloud Platforms: AWS (SageMaker, Glue, Lambda, S3), Azure ML, GCP
Big Data & ETL: Kafka, Airflow, ETL Pipelines, Databricks
Data Warehousing: BigQuery, Redshift, Snowflake
Data Modeling: Data warehouse design, Dimensional modeling, Star/Snowflake schema
Visualization Tools: Tableau, Power BI, Matplotlib, Seaborn
Containerization & Orchestration: Docker, Kubernetes
Statistics & Analytics: Descriptive statistics, Correlation, Hypothesis testing, A/B testing
Operating Systems: Windows, macOS, Linux

WORK EXPERIENCE

Data Engineer (AI/ML) **| Feb 2024 – Present**
McKinsey & Company, Dallas, TX

- Implemented Retrieval-Augmented Generation (RAG) models with probabilistic & statistical validation, reducing false positives in safety alerts by 35%.
- Developed and deployed NLP models that automated 85% of contract review processes, reducing manual effort and accelerating turnaround time by 40%.
- Built and optimized SQL and Python queries to extract large datasets for compliance and operational analysis, surfacing improved reporting accuracy.
- Partnered with cross-functional teams to design dashboards and automated reporting workflows in Tableau and Power BI for real-time performance tracking.
- Applied NLP and Generative AI techniques for text analytics, summarization, and synthetic data generation for automation-driven initiatives.
- Developed and productionized ML models using TensorFlow, PyTorch, Scikit-learn, and XGBoost, driving data-driven decision-making across business units.
- Designed and deployed scalable data pipelines using Azure Data Factory and Azure Synapse to automate ingestion from on-prem HR systems into Azure Data Lake (Gen2).
- Architect automated ETL workflows in AWS Glue, orchestrating data extraction, transformation, and loading processes to streamline data integration from various sources.
- Leverage the Boto3 library in Python to interact with AWS services like S3 and Lambda, automating cloud resource management and reducing manual intervention.
- Implement AI algorithms and models to solve complex problems, achieving a 25% improvement in accuracy and efficiency of data analysis.
- Utilized Apache Spark for large-scale data processing and Apache Kafka for real-time data streaming, ensuring efficient data movement and transformation.
- Created and maintained data warehouse structures using Snowflake, Redshift, and BigQuery, including dimensional modeling, star/snowflake schema design, and performance optimization.
- Leveraged BI tools like Tableau and Power BI for visualizing complex datasets, empowering business teams to make data-driven decisions, and contributing to an overall 15% increase in sales performance.

Data Engineer **| Jan 2023 - Feb 2024**
Morgan Stanley, Chicago, IL

- Applied advanced statistical techniques (hypothesis testing, regression, time-series analysis) to evaluate trading and risk datasets, improving model reliability by 25%.
- Automated complex data workflows using Apache Airflow, improving operational efficiency and reducing manual intervention by 50%, contributing to a more scalable, reliable data infrastructure.
- Performed detailed root cause analysis on abnormal trading system behaviors, enhancing risk-scoring accuracy by 22%.
- Engineered scalable SQL and Pandas pipelines to automate data validation for 10M+ high-frequency financial records daily, reducing manual oversight by 40% and improving data integrity.
- Streamlined monitoring workflows using Linux, Bash scripting, and Git, decreasing system downtime by 30% and boosting overall operational reliability.
- Engineered high-performance data transformation workflows on Databricks and Spark, optimizing computation workloads for multi-billion-row datasets.
- Developed a Redshift-based data delivery layer for business intelligence tools to operate directly on AWS S3.
- Crafted interactive data visualization dashboards using Tableau & Power BI, enabling stakeholders to derive actionable insights from complex datasets.
- Optimized SQL queries, Spark configurations, and cluster performance to reduce compute cost by 30% while improving job throughput.
- Contributed to the development, optimization, and maintenance of GCP workflows, leveraging BigQuery, Cloud Functions, and the creation of dashboards using Looker and Looker ML.
- Developed and deployed predictive forecasting models leveraging Databricks MLflow and Python, increasing forecasting accuracy by 18% and directly supporting improvements in product availability and revenue.
- Built end-to-end MLOps pipelines with MLflow for experiment tracking, versioning, and automated model deployment.

Data Analyst **| Aug 2020 - Aug 2022**
Mphasis, India

- Designed, built, and optimized ETL pipelines using AWS Glue and Lambda to process 10M+ financial transactions daily, ensuring high data reliability and scalability.
- Developed anomaly detection models and statistical validation workflows that enhanced fraud detection accuracy by 18% across financial datasets.
- Partnered with Data Science teams to automate model deployment, monitoring, and data validation for risk and compliance analytics, improving operational efficiency.
- Created interactive and automated Power BI and Tableau dashboards to monitor operational KPIs, data quality trends, and real-time performance metrics.
- Created and maintained interactive Excel VBA-based dashboards to facilitate data visualization and reporting for the site operations team
- Conducted A/B testing on digital engagement strategies, analyzing interactions from 50,000 users and supporting increased adoption of online tools
- Worked with ETL processes and optimized data pipelines for Power BI reporting and interacted with business stakeholders to gather requirements and convert them into actionable dashboards.
- Conducted thorough data cleaning and wrangling processes to enhance data quality and consistency across multiple projects, ensuring reliable analyses.

EDUCATION

Master Of Science - Management Information Systems
Southern Illinois University Edwardsville, Edwardsville, IL

CERTIFICATIONS

Statistics for Machine Learning & Deep Learning
AWS Certified Data Engineer – Associate
KPMG AU - Data Analytics Job Simulation

PROJECTS

Clinical Documentation Assistant – Self-Reflective RAG System - Link **| 2025**
Personal Project

- Developed an AI-assisted clinical documentation tool using Retrieval-Augmented Generation (RAG) to streamline healthcare workflows.
- Enabled smart summarization, ICD-10 code suggestions, and adaptive query refinement for enhanced documentation accuracy.
- Integrated ChromaDB, Sentence Transformers, and Streamlit, showcasing the real-world impact of ethical, transparent AI in clinical decision support.