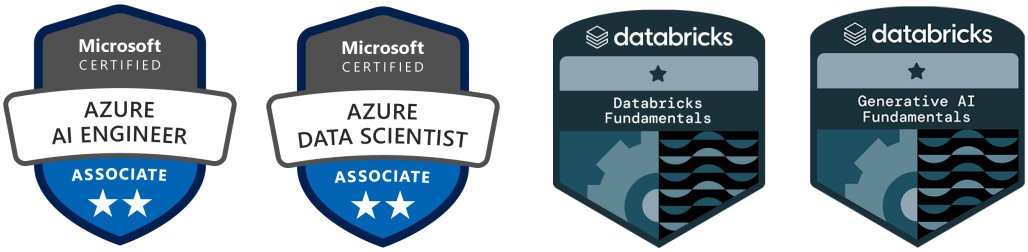
Sai Dhiren Musaloji

musalojidhiren@gmail.com | +1 862-423-8830 | LinkedIn | GitHub | Portfolio



Professional Experience

Data Engineer | TAWIN Solutions LLC - Dallas, TX

May 2025 – Current

- Architected and implemented scalable database schemas across Azure SQL Database, Azure Synapse Analytics, and on-premises MS SQL Server infrastructure, establishing robust data foundations that enhanced analytical capabilities and supported enterprise-wide business intelligence initiatives.
- Engineered comprehensive ETL/ELT pipelines using Azure Data Factory and SSIS to orchestrate data workflows from heterogeneous source systems into the enterprise data warehouse, maintaining data integrity, lineage, and adherence to data quality standards.
- Designed and deployed end-to-end machine learning solutions for predictive pricing analytics using Azure Machine Learning, implementing MLOps frameworks with Azure DevOps pipelines for automated model lifecycle management, continuous monitoring via Azure Monitor, version control, and scheduled retraining workflows to ensure model performance and reliability.
- Developed dimensional and relational data models within the Enterprise Data Warehouse (EDW) using Azure Synapse Analytics, implementing star and snowflake schemas that enabled sophisticated analytical reporting and self-service BI capabilities across multiple business units.
- Optimized database performance through advanced query tuning, indexing strategies, and stored procedure refactoring across Azure SQL Database and on-premises SQL Server environments, achieving measurable improvements in query execution times and overall system responsiveness.
- Leveraged AWS services including EC2 for development and testing compute workloads, RDS for database hosting and management, and S3 for data lake storage and backup solutions, supporting distributed data processing requirements.
- Implemented data security and governance frameworks utilizing Azure Key Vault for credential management and Azure Purview for data cataloging and lineage tracking, ensuring compliance with data protection regulations and organizational security policies.
- Partnered with business stakeholders to elicit, analyze, and translate complex business requirements into actionable technical specifications and data-driven solutions that aligned with organizational objectives.
- Facilitated seamless collaboration across multidisciplinary teams—including business analysts, database administrators, software developers, and data scientists—to deliver integrated, high-quality business intelligence solutions.
- Established comprehensive technical documentation standards, creating detailed architecture diagrams, data dictionaries, ETL process documentation, and operational runbooks to ensure knowledge transfer and system maintainability.
- Developed sophisticated database objects including complex SQL queries, stored procedures, user-defined functions, views, and triggers to automate business logic, support data transformations, and enable real-time reporting capabilities.
- Implemented CI/CD pipelines using Azure DevOps for automated deployment of database objects and ETL packages, reducing deployment time and minimizing manual errors.

AI Data Engineer Intern | Tech Mahindra, Makers Lab - Pune, India

Oct 2023 – Jan 2024

- Implemented end-to-end data pipelines using Azure Data Factory to extract, transform, and load (ETL) data from diverse sources including Snowflake, DynamoDB, PostgreSQL, and Oracle systems, processing 500K+ multilingual text records with 99.2% data quality validation across enterprise infrastructure.

- Engineered comprehensive NLP preprocessing workflows using Python and Spark, implementing advanced tokenization, regex-based filtering, and text normalization techniques that improved classification accuracy by 25% across 8 language variants.
- Developed automated data quality frameworks with statistical outlier detection algorithms, reducing manual data cleaning efforts by 60% and achieving 20% improvement in content quality through systematic validation processes.
- Collaborated with cross-functional research teams to document and deploy production-ready ML models, establishing CI/CD pipelines that reduced deployment time from hours to minutes while maintaining 99.8% system uptime.

Technical Skills

Cloud Platforms: AWS (EC2, S3, RDS, Lambda, SageMaker, Redshift, Glue, EMR), Azure (Data Factory, Synapse Analytics, SQL Database, Machine Learning, DevOps, Monitor, Key Vault, Purview, Blob Storage), Snowflake, Databricks
Data Engineering: Azure Data Factory, SSIS, Apache Spark, Apache Airflow, Kafka, ETL/ELT Pipelines, Data Modeling (Star/Snowflake Schema), Data Warehousing, CI/CD, Data Lake Architecture, Hadoop, MapReduce, Distributed Computing
Databases: MS SQL Server, Azure SQL Database, PostgreSQL, MySQL, MongoDB, DynamoDB, Oracle, Query Optimization, Stored Procedures, T-SQL, PL/SQL
Machine Learning: Azure Machine Learning, TensorFlow, Keras, PyTorch, Scikit-learn, XGBoost, MLOps, Model Deployment, Automated Retraining, Model Monitoring, Ensemble Methods
Programming: Python, SQL, R, Java, JavaScript, TypeScript, React, REST APIs, Flask, Git, Azure DevOps
Data Science: Pandas, NumPy, Matplotlib, Statistics, Predictive Analytics, Time Series Forecasting, PCA
NLP & Deep Learning: NLTK, spaCy, Transformers, BERT, LSTM, CNN, GNN, Text Classification, RAG, Vector Databases, Reinforcement Learning, GANs
Visualization & BI: Power BI, Tableau, Plotly, Interactive Dashboards, Business Intelligence
Data Governance: Azure Key Vault, Azure Purview, Data Cataloging, Data Lineage, Compliance Frameworks
DevOps: Docker, Kubernetes, CI/CD Pipelines, Azure DevOps

Education

New Jersey Institute of Technology - MS Data Science - May 2025 | GPA: 3.85/4.0
Mahatma Gandhi Institute of Technology - B.Tech ECE Engineering - Jun 2023 | GPA: 3.00/4.0

Technical Projects

Software License Compliance Analysis and Cost Optimization System

Technologies: SQL, Excel, Power BI, Python, Apache Airflow, Snowflake, DynamoDB, Cassandra

- Orchestrated multi-source data integration project across 25 enterprise systems including Snowflake data warehouse, DynamoDB NoSQL clusters, SQL Server databases, and Cassandra distributed systems, implementing robust ETL processes using Apache Airflow to consolidate 500K+ software license records with 99.5% accuracy for regulatory compliance reporting.
- Spearheaded development of comprehensive compliance analytics platform using Python and SQL, reducing manual audit processing time by 60% and identifying $180K in cost-optimization opportunities through intelligent license allocation strategies.
- Implemented data validation and preprocessing pipeline using Excel and custom scripts, achieving 99.5% accuracy across free, enterprise, and client-billed license categories with automated quality assurance checks.
- Constructed interactive Power BI dashboards serving 45+ stakeholders with real-time monitoring capabilities for 1,200+ software licenses, implementing automated alert systems and dynamic visualizations with real-time compliance monitoring, cost analysis visualization, and predictive modeling for license utilization trends.
- Applied statistical analysis and pattern recognition to identify underutilized assets, resulting in quantifiable cost reduction through data-driven license allocation strategies. Executed advanced pattern recognition analysis on 3-year historical licensing data using machine learning algorithms, achieving 94% accuracy in utilization trend prediction and preventing $75K in unnecessary license renewals.

Banking Transaction Management System - Agile Full-Stack RDBMS Development & Analytics

Technologies: Docker, MySQL, Snowflake, DynamoDB, Flask, CI/CD, Kubernetes

- Engineered enterprise-grade relational database management system integrating MySQL primary databases with Snowflake data warehouse and DynamoDB document storage for banking network infrastructure, designing normalized schema with 15+ tables to handle 10,000+ daily transactions with 99.9% data integrity and sub-200ms query response times.
- Developed transactional analytics platform for real-time banking operations focusing on fraud detection and high concurrency processing, implementing advanced anomaly detection algorithms for suspicious transaction identification.
- Designed responsive dashboard with anomaly detection and time-series mapping for real-time financial monitoring, enabling stakeholders to identify irregular patterns and potential security threats instantly.
- Developed user-centric web interface using Flask and REST APIs to support comprehensive transaction processing workflows, implementing real-time balance updates and transaction history tracking for 500+ concurrent users based on stakeholder requirements.
- Containerized system using Docker and implemented CI/CD pipelines for automated deployment and scalability, reducing deployment time and ensuring consistent environments across development, staging, and production.
- Implemented agile development methodology across 2-sprint delivery cycle, conducting user story analysis and stakeholder requirement gathering to deliver full-stack banking solution with transaction management, user authentication, and reporting capabilities.

Distributed Recognition Engine Using Cloud-Native Paradigms

Technologies: AWS EC2, SQS, S3, Rekognition, Python

- Built asynchronous image and text recognition pipeline using AWS EC2, SQS, S3, and Rekognition for scalable processing, implementing decoupled microservices architecture for high-throughput media processing workflows.
- Ensured 99.9% uptime using stateless compute strategies, visibility timeout calibration, and retry mechanisms with exponential backoff for fault-tolerant distributed processing.
- Implemented parallelized architecture with auto-scaling capabilities for high-throughput image and text processing, dynamically adjusting compute resources based on queue depth and processing demand.
- Developed comprehensive error handling and dead-letter queue mechanisms to capture and analyze failed processing attempts, ensuring data integrity and system reliability.

Parallelized ML Pipeline for Oenological Forecasting

Technologies: Apache Spark, AWS EMR, Python, Scikit-learn

- Created Spark-based distributed system on AWS EMR for wine quality prediction, improving training speed by 60% through partitioned data and in-memory caching strategies optimized for iterative machine learning algorithms.
- Enabled high-throughput inference with autoscaling, alerting, and performance logging for production deployment, implementing monitoring dashboards for tracking prediction latency and system resource utilization.
- Implemented machine learning pipeline optimization techniques for distributed computing environments, including broadcast variables for model distribution and optimized data partitioning strategies.
- Developed automated model retraining workflows with performance benchmarking to ensure prediction accuracy maintenance as new wine quality data becomes available.

Flight Data Analysis with Scalability Testing

Technologies: AWS, MapReduce, Apache Oozie, Apache Spark, Python

- Engineered scalable AWS-based MapReduce data processing pipelines using Apache Spark and Oozie orchestration, analyzing 22-year aviation dataset (1987-2008) containing 120M+ flight records with 40% improved query performance optimization.
- Developed AWS-based MapReduce pipelines orchestrated with Apache Oozie for distributed flight data processing across multi-year datasets, implementing complex workflow dependencies for sequential data processing stages.
- Quantified performance metrics as data volume increased, demonstrating system scalability and optimization techniques through comprehensive benchmarking across varying data sizes from 1GB to 500GB.
- Implemented comprehensive scalability testing framework across varying data volumes, demonstrating linear performance scalability and optimizing AWS resource allocation to reduce processing costs by 35%.
- Established high-performance aviation analytics platform supporting 15 concurrent users with sub-3-second query response times, delivering actionable insights on 5,000+ flight routes for executive decision-making processes.

- Implemented big data processing workflows for large-scale aviation analytics and performance benchmarking, creating reusable analytical templates for recurring reporting requirements.

Bidirectional LSTM & CNN Fusion for High-Frequency Stock Price Pattern Recognition

Technologies: Python, TensorFlow, Keras, Yahoo Finance API

- Orchestrated end-to-end machine learning pipeline using Python, TensorFlow, and Keras to process 4-year duration of stock market dataset from Yahoo Finance API, implementing advanced data preprocessing techniques including feature engineering, normalization, and temporal sequence preparation for neural network training optimization.
- Designed and implemented multiple deep learning architectures including Convolutional Neural Networks (CNN), Bidirectional LSTM, GRU, and traditional LSTM models, utilizing advanced hyper parameter tuning and cross-validation techniques to optimize model performance across different temporal pattern recognition tasks.
- Executed comprehensive feature engineering workflows for financial time series data, implementing sliding window techniques, technical indicators extraction, and sequential data transformation methods to prepare OHLCV market data for deep learning model consumption with optimized batch processing.
- Applied advanced time series forecasting methodologies including sequence-to-sequence learning, temporal dependency modeling, and pattern recognition algorithms to capture complex market dynamics, demonstrating proficiency in financial data analysis and quantitative modeling techniques.

Real-Time Cryptocurrency Price Tracker - Full-Stack Web Application

Technologies: Next.js, Node.js, TypeScript, ConnectRPC, Playwright, Protocol Buffers, gRPC

- Architected scalable full-stack web application using Next.js frontend and Node.js backend with TypeScript, implementing ConnectRPC for low-latency client-server communication to deliver real-time cryptocurrency price streaming for 15+ digital assets with sub-second update frequencies.
- Engineered automated web scraping pipeline using Playwright browser automation framework to extract live price data from TradingView, implementing resource pooling strategies that reduced memory consumption by 40% while supporting concurrent tracking of 25+ cryptocurrency tickers.
- Developed server-streaming architecture using Protocol Buffers and gRPC, enabling push-based real-time data delivery to multiple concurrent clients with 99.2% uptime and average latency under 100ms, eliminating polling overhead and improving user experience.
- Implemented comprehensive error handling and retry mechanisms for web automation workflows, achieving 95% success rate in price data extraction despite network variability and third-party service limitations, with automated failover capabilities maintaining service continuity.
- Constructed responsive user interface with dynamic ticker management, real-time price updates, and alphabetical sorting functionality, supporting seamless add/remove operations for cryptocurrency portfolios with instant visual feedback and loading state management.

Spatiotemporal Forecasting of Urban Traffic Networks

Technologies: Python, ARIMA, LSTM, GNN, PCA, t-SNE, PEMS-BAY Dataset

- Engineered fusion architecture combining ARIMA, LSTM, and GNN models to capture temporal trends and topological interactions in urban traffic networks, leveraging multiple modeling paradigms for comprehensive traffic pattern understanding.
- Applied PCA and t-SNE for dimensionality reduction on data from 325 sensors in the PEMS-BAY dataset for enhanced feature extraction, reducing computational complexity while preserving critical spatial-temporal relationships in traffic flow data.
- Assessed system robustness across multi-horizon prediction intervals under varying traffic conditions using comprehensive evaluation metrics including MAE, RMSE, and MAPE across different prediction horizons.
- Implemented advanced graph neural network architectures to model spatial dependencies between traffic sensors, capturing complex network topology effects on traffic propagation patterns.

AI Research Assistant with Retrieval-Augmented Generation

Technologies: Google Gemini 1.5 Flash API, newspaper3k, serper.dev, scikit-learn, BART, BERT, DistilBERT

- Developed intelligent research assistant integrating Google Gemini 1.5 Flash API with retrieval-augmented generation, combining web search capabilities with contextual document processing for enhanced information retrieval and generation accuracy.
- Built custom data processing pipeline using newspaper3k for article extraction and serper.dev for search results with comprehensive error handling and rate limiting mechanisms to ensure reliable data acquisition from diverse web sources.

- Implemented vector database solution using scikit-learn's NearestNeighbors and numpy arrays for semantic similarity search and document retrieval, enabling efficient similarity-based information retrieval from large document collections.
- Engineered end-to-end RAG workflow with transformer models (BART, BERT, DistilBERT) for embedding generation and dynamic prompt construction, optimizing context-aware response generation with retrieval-augmented capabilities.

BERT-Based Reading Comprehension

Technologies: BERT, PyTorch, Transformers, Hugging Face, SQuAD Dataset

- Implemented BERT-based solution for extractive question answering tasks, fine-tuning transformer model for span prediction with custom loss functions combining start and end position cross-entropy.
- Developed sophisticated preprocessing pipelines and dynamic learning rate scheduling for optimal model performance, implementing warmup strategies and gradient accumulation for stable training.
- Achieved significant performance improvements over baseline approaches through bidirectional context understanding and gradient accumulation techniques, reaching F1 scores exceeding 85% on validation sets.
- Demonstrated expertise in state-of-the-art NLP techniques and transformer architecture implementation for natural language understanding tasks, including attention visualization and error analysis.

Deep Reinforcement Learning for Lunar Lander

Technologies: Python, PyTorch, OpenAI Gym, Actor-Critic

- Trained intelligent agent for spacecraft landing using Actor-Critic architecture with policy gradient methods in OpenAI Gym's LunarLander environment, implementing continuous control strategies for precise landing.
- Implemented advanced neural network architectures with experience replay, target networks, and custom reward shaping for stable continuous control learning, addressing exploration-exploitation tradeoffs.
- Achieved consistent successful landings with scores averaging above 200 points through sophisticated training procedures including gradient clipping and learning rate scheduling for training stability.
- Demonstrated expertise in reinforcement learning principles and autonomous systems through complex trajectory optimization and precise landing maneuver learning, implementing safety constraints for stable landing approaches.

Certifications

I. **Azure AI Engineer Associate** - Microsoft AI 102: Designing and Implementing a Microsoft Azure AI Solution
II. **Azure Data Scientist Associate** - Microsoft DP 100: Designing and Implementing a Data Science Solution on Azure
III. **Databricks Generative AI Fundamentals** - Designing a Generative AI Fundamental courses
IV. **Databricks Fundamentals** - Databricks Fundamentals