Jay Singhvi

2066603609 | jay.singhvi@outlook.com | Portfolio: jay-singhvi.github.io/ | linkedin.com/in/jay-singhvi/ | github.com/jay-singhvi/

WORK EXPERIENCE

Data Scientist (RA) — **Seattle University**, Seattle, WA — Sep 2022 – Ongoing

- Built ***HIPAA-compliant data pipelines*** using ***Docker*** and ***AWS*** to process medical datasets for asthma research, implementing ***transfer learning*** models that achieved ***88% accuracy*** in predicting asthma onset - ***20% better*** than traditional methods
- Developed ***ensemble ML models*** using ***PySpark*** and ***Python*** for personalized healthcare predictions, improving forecasting accuracy by ***12%*** over standard neural networks while handling sparse medical data

Data Engineer — **Yardi Systems**, Dubai, UAE — Apr 2019 - Jul 2022

- Designed and deployed ***ETL pipelines*** using ***SSIS*** and ***SQL Server*** for real estate BI modules, reducing deployment costs and implementation time by ***75%*** through automated data extraction and transformation
- Built ***interactive BI dashboards*** with real-time data streaming that reduced data inconsistencies by ***90%***, for better decision-making
- Created scalable data ingestion systems supporting multiple data formats (batch, real-time, historical), accelerating customer onboarding by ***60%*** and improving processing speed by ***40%*** through ***SQL Server partitioning***
- Automated reporting solutions using ***T-SQL stored procedures***, triggers, and optimized indexes, saving ***20 hours weekly*** and improving query performance through execution plan optimization

Data Engineer — **Yardi Systems**, Pune, India — Nov 2016 - Mar 2019

- Built ***ETL systems*** using ***SSIS*** and Yardi frameworks to streamline lease approval workflows, improving system usage by ***50%*** and securing ***$3M+ revenue retention***
- Designed ***data warehouse*** with dimensional modeling to power real-time dashboards tracking ***100+ KPIs*** for property management, reducing processing time by ***15%*** through incremental data loading
- Developed ***Data Mart solutions*** with user-friendly interfaces for complex property datasets, establishing data governance standards with automated testing and documentation
- Optimized database performance through strategic indexing and query tuning for large-scale environments, while building integration frameworks using ***SSDT*** to process multiple data sources (Excel, flat files, databases)

EDUCATION

MS, Computer Science (specialization in Data Science) — **Seattle University**, Seattle, WA — Sept 2022 – Jun 2024

MS, Computer Applications — **Symbiosis International University**, India — July 2015 - Apr 2018

BS, Information Technology — **University of Mumbai**, Mumbai, India — Jun 2011 – Jan 2015

PUBLICATIONS & CERTIFICATIONS (*Research Papers:* github.com/jay-singhvi/publications)

- Published in DAWAK 2024 - **Incremental SMOTE with Control Coefficient for Classifiers in Data Starved Medical Applications**
- Published in SAC_2025 - **A Retrieval-Augmented Framework for Meeting Insight Extraction**
- Peer-review for IEEE JBHI 2025 - **Hybrid Deep Learning using Transfer Learning as Feature Extractor in Env. Health Risk Prediction**
- CITI Program - Responsible Conduct of Research – Engineers | Human Subjects Research for IRB (Faculty, Staff, and Student) *(Other Certificates:* **linkedin.com/in/jay-singhvi/details/certifications/***)*

PROJECTS *(GitHub Portfolio*: github.com/jay-singhvi/*)*

Resonate AI Chatbot: *(Tech Stack: Python, Transformers, LangChain, Pinecone, Hugging Face, LLM, RAG, AWS S3 & AWS Transcribe, QLoRA)*

- Built ***RAG system*** using ***LangChain*** with semantic graph clustering, achieving ***90% BERT similarity scores*** and ***89% precision/recall*** through overlapping document chunking that preserved context across boundaries
- Developed high-performance vector embedding layer using ***FAISS*** (local) and ***Pinecone*** (cloud) that maintained ***85% cosine similarity*** while optimizing dimensional reduction for fast query performance in high-volume scenarios
- Created comprehensive ***LLM evaluation framework*** across ***OpenAI***, ***Anthropic***, and ***Google*** models, tracking hallucination rates and response accuracy while building semantic routing that ***reduced costs*** by selecting optimal models based on query type
- Fine-tuned ***Llama 2 (7B)*** using ***QLoRA techniques***, reducing computational requirements by ***70%*** through gradient checkpointing and mixed precision training while building specialized datasets for enterprise use cases
- Built distributed inference system with intelligent caching that reduced response latency by ***65%*** while maintaining quality, establishing ***CI/CD pipeline*** for continuous improvements

AI-Agentic Synthetic Data Generation: *(Tech Stack: Python, Docker, Anthropic API, Claude AI, CSV manipulation, CLI)*

- Built containerized ***AI system*** with specialized agent architecture featuring analyzer and generator components that produce synthetic datasets with statistical fidelity to source distributions through modular framework design
- Engineered error handling with comprehensive logging and ***batch processing framework*** that dynamically adjusts parameters based on memory and CPU usage, optimizing throughput through parallelized pipelines
- Leveraged ***Anthropic Claude 3.5 Sonnet*** through context-aware prompt engineering that preserved statistical properties, implementing adaptive prompting strategies and validation pipelines for synthetic data verification
- Developed parameter validation, contextual help systems, and secure ***API key management*** with environment-based configuration
- Published the containerized solution to ***Docker Hub*** and versioned releases for widespread adoption

Personalized Marketing Campaign Optimizer: *(Tech Stack: Python, Scikit-learn, Pandas, Matplotlib, Seaborn, SMOTE, GridSearchCV)*

- ***Architected*** marketing optimization system processing ***500,000+ customer records*** using ensemble ***ML techniques*** (***Decision Tree, KNN, Random Forest***), achieving ***86% prediction accuracy*** and identifying high-conversion segments representing ***$2M+ revenue impact***
- ***Engineered*** advanced feature engineering pipelines transforming ***50+ raw data sources*** into predictive indicators, implementing ***SMOTE*** and ***Random Under Sampling*** techniques that improved minority class prediction by ***25%*** without accuracy loss
- ***Designed*** robust ***EDA workflows*** with custom visualizations processing ***10GB+ datasets***, revealing previously undetected patterns across ***15+ market segments*** and enabling ***30% faster model deployment***
- ***Implemented*** automated data quality systems with statistical outlier detection processing ***1M+ records daily***, substantially improving pipeline integrity and reducing manual validation time by ***80%***
- ***Optimized*** model performance using ***stratified cross-validation*** and ***GridSearchCV*** across ***100+ hyperparameter combinations***, establishing automated workflows achieving ***15% improvement*** in production model accuracy

SQL Query Assistant using Snowflake Cortex Analyst: *(Tech Stack: AWS S3, Python, Snowflake, Streamlit, SQL, LLM, Snowflake Cortex LLM)*

- ***Architected*** advanced ***NLP-to-SQL*** system using ***Snowflake Cortex Analyst*** processing ***5,000+ daily queries***, transforming natural language into optimized SQL with ***95% accuracy*** for non-technical users
- ***Engineered*** comprehensive semantic model framework with ***YAML configurations*** defining ***200+ logical tables*** and ***500+ measures***, enabling high-precision query generation across ***multi-TB data warehouses***
- ***Developed*** production-grade ***Streamlit chatbot*** handling ***1,000+ concurrent users*** with robust error handling and session state management, reducing query composition time by ***70%*** for business analysts
- ***Implemented*** optimized ***ETL pipelines*** ingesting ***multiple revenue datasets*** (***100GB+ daily volume***) with precise data type handling, ensuring ***99.99% data integrity*** across analytical ecosystem
- ***Designed*** intelligent caching mechanisms and verified query repository capturing ***10,000+ validated SQL patterns***, reducing ***API call frequency by 60%*** and improving system response time to ***<500ms***

Asthma Patient Research Project *(South Korean Hospital Collaboration):* *(Tech Stack: Python, Scikit-learn, Pandas, NumPy, SciPy, Matplotlib, Seaborn, OpenWeatherMap API, Jupiter Notebooks, PostgreSQL)*

- Spearheaded clustering analysis on patient records implementing diverse algorithms (K-means, DBSCAN, Affinity Propagation, BIRCH, Mean-Shift, OPTICS), integrating weather and air quality datasets from 10+ external APIs (OpenWeatherMap, EPA AQS, Ambee) for environmental analysis
- Implemented Lag-Llama foundation model within customized Python prediction framework processing time-series data (1M+ data points), conducting rigorous comparative analysis using statistical metrics (MSE, RMSE, MAE, R^2, MAPE) and cross-validation techniques
- ***Performed*** systematic hyperparameter optimization using ***GridSearchCV and RandomSearchCV*** expanding usable training data segments by ***37%*** while maximizing prediction accuracy on ***missing time-series values***, collaborating with ***research team of 8+ medical professionals*** through ***PostgreSQL database integration***

Agricultural Computer Vision Project *(Washington State Farmers Collaboration):* *(Tech Stack: Python, OpenCV, PyTorch, Ultralytics YOLO, NumPy, Pandas, Matplotlib, PIL/Pillow, Open3D, ML-Depth-Pro, CVAT, Roboflow, scikit-image, DroneKit, Jupyter Notebooks)*

- Engineered 3D visualization pipeline using Open3D and OpenCV transforming 2,000+ drone images (50GB+ imagery data) into comprehensive volumetric models, implementing Ultralytics YOLOv10 architecture achieving 89.8% accuracy in automated plant counting
- Orchestrated aerial data collection strategy processing 7,000+ unique bounding boxes using CVAT and Roboflow annotation platforms, establishing PyTorch-based ML data pipeline supporting real-time processing of 100+ images/hour with OpenCV preprocessing
- Conducted comparative performance analysis on YOLO architectures (v8, v11, v12) using PyTorch and Ultralytics framework, achieving 93.6% accuracy with YOLOv12 and integrating ML-Depth-Pro libraries and OpenCV for distance measurement and size estimation

TECHNICAL SKILLS

- **Cloud & Infrastructure**: AWS (S3, EC2, EKS, Lambda, IAM, DynamoDB, SageMaker, Transcribe, CloudFormation, SDK, CLI), Docker, Git, CI/CD Pipelines (GitHub Actions), Team Foundation Version Control (TFS)
- **Data Engineering & ETL**: SQL Server Integration Services (SSIS), ETL Pipeline Development, Data Modeling (Fact/Dimension), Incremental Loading, Parallel Processing, Error Handling, Data Pipeline Monitoring, Automated Data Upload, System Analysis & Design, SDLC
- **Databases & Query Languages**: SQL Server, PostgreSQL, MySQL, NoSQL, T-SQL, Stored Procedures, Triggers, User Defined Functions (UDF), Views, Query Optimization, Database Partitioning, Index Management
- **Big Data & Analytics Platforms**: Snowflake, Snowflake Cortex Analyst, Data Warehousing, Data Marts, OLAP Cubes
- **Machine Learning & AI**: Machine Learning (Supervised/Unsupervised), Deep Learning, Transfer Learning, Ensemble Modeling, Neural Networks, Feature Engineering, A/B Testing, LLMs (OpenAI, Claude, Google), RAG, Random Forest, Decision Trees, GridSearchCV, Cross-validation, Hyperparameter Tuning
- **Programming & Development**: Python, PySpark, RESTful APIs, Bash Scripting, API Development, Object-Oriented Programming, Environment Management, CLI Development, Batch Processing, Docker Compose
- **Data Science Libraries & Tools**: LangChain, Pinecone, NumPy, Pandas, PyTorch, Scikit-learn, TensorFlow, Streamlit, Spark SQL, Google Colab, Jupyter Notebook, Hugging Face
- **Visualization & Reporting**: Matplotlib, Seaborn, Power BI, Tableau, Custom Dashboards, KPI Tracking, Real-time Data Visualization, Automated Reporting Solutions, Business Metrics Tracking