# Jay Singhvi

2066603609 | jsinghvi@seattleu.edu | Portfolio | linkedin.com/in/jay-singhvi/ | github.com/jay-singhvi/

Data Engineer with 7+ years of experience specializing in CRM platforms, cloud migrations, and data pipeline development. Proven track record in modernizing legacy systems, implementing ETL solutions, and developing scalable data architectures. Expertise in SQL, database optimization, and cloud technologies with a focus on delivering high-impact operational and analytical solutions.

## TECHNICAL SKILLS

- **Cloud & Infrastructure**: AWS (S3, EC2, EKS, DynamoDB, Transcribe, SageMaker, SDK, CLI), Docker, Kubernetes, Team Foundation Version Control (TFS), Git, CI/CD Pipelines
- **Data Engineering & ETL**: SQL Server Integration Services (SSIS), ETL Pipeline Development, Data Modeling (Fact/Dimension), Data Validation, Incremental Loading, Parallel Processing, Error Handling, Data Quality Control, Data Governance, Version Control, Data Pipeline Monitoring, Automated Data Upload, System Analysis & Design, SDLC
- **Databases & Query Languages**: SQL Server, MySQL, NoSQL, DynamoDB, T-SQL, Stored Procedures, Triggers, User Defined Functions (UDF), Views, Query Optimization, Database Partitioning, Index Management, SQL Server Agent
- **Big Data & Analytics Platforms**: Hadoop, MapReduce, Snowflake, Snowflake Cortex LLM, Business Intelligence Tools, Data Warehousing, Data Marts, OLAP Cubes, Data Warehouse
- **Machine Learning & AI**: Machine Learning (Supervised/Unsupervised), Deep Learning, Transfer Learning, Ensemble Modeling, SMOTE, Neural Networks, Feature Engineering, A/B Testing, LLMs (OpenAI, Claude, Google), RAG, QLoRA, BERT, Semantic Clustering, Random Forest, KNN, Decision Trees, GridSearchCV, Cross-validation, Hyperparameter Tuning
- **Programming & Development**: Python, PySpark, Flask, RESTful APIs, OAuth 2.0, Bash Scripting, API Development, Object-Oriented Programming, Environment Management, CLI Development, Batch Processing
- **Data Science Libraries & Tools**: LangChain, Pinecone, NumPy, Pandas, PyTorch, Scikit-learn, TensorFlow, Streamlit, Spark SQL, Google Colab, Jupyter Notebook, Hugging Face, Transformers, Natural Language Processing, Semantic Routing, Prompt Engineering
- **Visualization & Reporting**: Matplotlib, Seaborn, Power BI, Custom Dashboards, KPI Tracking, Real-time Data Visualization, Ad-hoc Reporting Tools, Automated Reporting Solutions, Business Metrics Tracking

## WORK EXPERIENCE

**Research Assistant (Data Science)** — **Seattle University**, Seattle, WA — Jun 2023 – Ongoing

- Conducted HIPAA-compliant research for asthma patients using transfer learning techniques to develop personalized ensemble models.
- Achieved 88% accuracy in asthma prediction by performing comparative analysis while working with sparse medical data. This represented a 20% improvement compared to existing classifiers and a 12% improvement over neural network models.
- Completed CITI Program certifications in research ethics and human subject research, ensuring compliance and ethical conduct in data handling.
- Published research papers outlining findings and detailing advanced machine learning techniques for enhancing medical predictions.
- Currently working on research projects with the Korean government on historic pattern analysis of asthma patients and with Washington state farmers for farming yield prediction using UAV and IOT for accurate insurance premiums and smoother claims process.

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

- Designed and implemented comprehensive ETL pipelines using SSIS for data extraction from multiple source systems, performing complex transformations, and loading into enterprise data warehouse, resulting in 75% reduction in deployment costs and time for Yardi's real estate business intelligence module.
- Developed and maintained BI dashboards using in-house Tool (similar to Power BI), incorporating real-time data updates and custom visualizations for business metrics tracking and reporting, improving data visibility and decision-making processes.
- Orchestrated cross-functional collaboration across 4 teams spanning 3 time zones, successfully managing BI transformation projects for 50+ Middle East clients, ensuring seamless data integration and maintaining 99.99% system availability.
- Implemented robust data quality checks and validation procedures throughout the ETL process, reducing data inconsistencies by 90% and improving overall data accuracy.
- Created scalable data ingestion frameworks supporting multiple data types (historical, event-based, batch) and sources, reducing customer onboarding time by 60% through automated data upload processes.
- Optimized ETL workflows through parallel processing and incremental load strategies, achieving 40% improvement in data processing times while maintaining data integrity.
- Implemented monitoring and alerting systems for ETL jobs, reducing system downtime by 85% through proactive issue identification and resolution.
- Developed comprehensive documentation for ETL processes, data models, and BI dashboards, facilitating knowledge transfer and reducing training time for new team members by 50%.

- Established data governance protocols and security measures ensuring compliance with industry standards while maintaining data accessibility for authorized users.
- Created automated reporting solutions for tracking KPIs and business metrics, saving approximately 20 hours per week in manual reporting efforts.

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

- Designed and implemented complex ETL workflows using Yardi's proprietary tools and SQL Server Integration Services (SSIS) to streamline lease approval processes, resulting in 50% increased system utilization and $3M+ revenue retention.
- Developed and optimized data warehousing solutions including fact tables, dimension modeling, and ETL packages to create real-time dashboards tracking 100+ KPIs across property management metrics.
- Trained 4 new engineers in implementing best practices for data pipeline development, code review, and deployment processes.
- Created and maintained stored procedures, triggers, and complex SQL queries to optimize data extraction and transformation processes for rental agreement automation.
- Implemented error handling and logging mechanisms in ETL packages to ensure data quality and maintain system reliability.
- Developed automated testing frameworks for data validation and performance monitoring of ETL processes.
- Designed and implemented incremental load strategies for large-scale data processing, reducing processing time by 15%.
- Established data governance protocols and documentation standards for ETL processes and warehouse structures.
- Created user-friendly interfaces and reporting tools for stakeholders to access and analyze property management data.
- Responsible for Deploying, Scheduling Jobs, Alerting and Maintaining SSIS packages using SQL Server Agent, and managing version control through Team Foundation Version Control (TFS).
- Optimized database performance through index management, query tuning, and implementation of partitioning strategies.

**Junior Data Engineer** **Yardi Systems**, Pune, India Feb 2015 - Oct 2016

- Designed and developed various SSIS packages (ETL) to extract and transform data and involved in Scheduling SSIS Packages.
- Employed Data warehousing techniques to develop a comprehensive Data Mart, serving as a reliable data source for downstream reporting. Developed a User Access Tool empowering users to create ad-hoc reports and execute queries for in-depth analysis within the proposed Cube.
- Worked extensively on system analysis, design, development, testing, and implementation of projects (Complete SDLC)
- Performed SSIS Development and support, developed ETL solutions for integrating data from multiple sources like Flat Files (delimited, fixed width), Excel, SQL Server, and Raw File using SQL Server Data Tools (SSDT)
- Identify and resolve problems encountered during both the development and release of the SSIS code.
- Debugging and troubleshooting technical issues while implementing the applications.
- Created Complex ETL Packages using SSIS to extract data from staging tables to partitioned tables with incremental load utilizing SQL Server partitioning strategies.
- Created packages in SSIS with error handling and worked with different methods of logging in SSIS.
- Created complex Stored Procedures, Triggers, Functions (UDF), Indexes, Tables, Views, and other T-SQL code and SQL joins for applications following SQL code standards.
- Performed efficient tuning of SQL source queries for data load using SQL Server query optimization techniques and execution plans.
- Created views to display required information on the user interface, and triggers to validate consistent data entry into the Microsoft SQL Server database.

## EDUCATION

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

*Recipient of Seattle University's Dean's Honor Roll*

*Courses: Distributed Systems, Machine Learning, Big Data Analytics, AWS Cloud Computing, Artificial Intelligence*

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

*Courses: Python, Linux scripting, Data Structure Algorithms, Relational Database management, Data Mining and Warehousing*

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

*Courses: Database management, SQL, Linux, Data Warehousing*

## PUBLICATIONS & CERTIFICATIONS

- **Incremental SMOTE with Control Coefficient for Classifiers in Data Starved Medical Applications**, published in the 26th International Conference on Big Data Analytics and Knowledge Discovery (DAWAK 2024).
- **A Retrieval-Augmented Framework for Meeting Insight Extraction**, accepted to be published in SAC_2025 (The 40th ACM/SIGAPP Symposium on Applied Computing, Track: Intelligent Systems for Digital Era)
- **Hybrid Deep Learning Framework using Transfer Learning as Feature Extractor in Env. Health Risk Prediction**, in peer-review IEEE JBHI 2025
- CITI Program - Responsible Conduct of Research – Engineers | Human Subjects Research for IRB (Faculty, Staff, and Student) and other **certificates**

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

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

- Developed RAG system for meeting insight extraction: Optimized with semantic graph clustering (90% BERT similarity, 89% precision/recall); stored embeddings in Pinecone (85% cosine similarity).
- Improved response quality: A/B tested prompts and LLMs (OpenAI, Claude, Google); implemented semantic routing; fine-tuned Llama 2 (7B) with QLoRA for enhanced chatbot persona.
- Deployed Streamlit prototype on Hugging Face Spaces, contributing to open-source AI community.

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

- Developed Docker-containerized AI agent with Claude AI: Analyzer Agent for CSV analysis and Generator Agent for synthetic data production; implemented batch processing for efficient large dataset generation.
- Leveraged Anthropic API and Claude 3.5 Sonnet: Used advanced prompt engineering for high-fidelity data generation; designed user-friendly CLI for input specification and output customization.
- Enhanced usability and distribution: Implemented robust CSV handling; secured API key management; published Docker image to Docker Hub; created comprehensive documentation for open-source community.

**Serverless Employee Management System:** *(Tech Stack: Python, AWS S3 & AWS DynamoDB, Docker, AWS EKS, RESTful APIs, OAuth 2.0, Microservices)*

- Built cloud-native SAAS for workforce management: Used Python, AWS (S3, DynamoDB); implemented microservices architecture; developed RESTful APIs with OAuth 2.0 for security and integration.
- Optimized deployment and scaling: Utilized Docker for containerization, Kubernetes for orchestration; implemented serverless architecture to reduce costs and improve resource utilization.

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

- Developed marketing campaign optimizer: Used Python, Scikit-learn; implemented Decision Tree, KNN, and Random Forest classifiers; achieved 86% accuracy; addressed class imbalance with SMOTE and Random Under Sampling.
- Enhanced data pipeline and analysis: Engineered features for improved performance; conducted EDA using Pandas, Matplotlib, and Seaborn; implemented cross-validation and hyperparameter tuning for optimized model performance across customer segments.

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

- Created SQL query creator based on Natural Language input from Streamlit chatbot, helps people with No SQL knowledge to create and execute queries.

## ADDITIONAL PROJECTS

- **2048 AI Player**: Created an Autonomous player for the game 2048 using Python, employing multiple search algorithms for decision-making, and predicting the best move.
- **ImgProcessor**: Designed and implemented a website for image processing and manipulation using Python and Flask on RPC.
- **H2OQualitizer**: Built a water quality prediction system using Python. Explored various machine learning models like regression, classification, and clustering. Processed data by handling outliers and missing values. Achieved 66.8% accuracy with the MLP Classifier model. Utilized Git for version control and collaboration.