Jay Singhvi

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

Data Engineer with 9+ years of experience architecting high-performance data ecosystems across cloud platforms and ETL frameworks. Delivered measurable business impact: 75% reduction in deployment costs, 90% decrease in data inconsistencies, and 60% faster customer onboarding while maintaining 99.99% system uptime. MS in Computer Science with published research in healthcare analytics. Expertise in RAG frameworks, containerized AI solutions, serverless architectures, and advanced data engineering across AWS, Snowflake, Python, and SQL.

TECHNICAL SKILLS

- **Cloud & Infrastructure**: AWS (S3, EC2, EKS, Lambda, Step Functions, IAM, DynamoDB, SageMaker, Transcribe, CloudFormation, SDK, CLI), Terraform, Docker, Kubernetes, Git, CI/CD Pipelines (GitHub Actions), Team Foundation Version Control (TFS)
- **Data Engineering & ETL**: Databricks, Delta Lake, dbt, Apache Airflow, SQL Server Integration Services (SSIS), ETL Pipeline Development, Data Modeling (Fact/Dimension), Change Data Capture (CDC), Incremental Loading, Parallel Processing, Error Handling, Data Pipeline Monitoring, Automated Data Upload, System Analysis & Design, SDLC
- **Databases & Query Languages**: SQL Server, PostgreSQL, MySQL, NoSQL (MongoDB, DynamoDB), GraphQL, T-SQL, Stored Procedures, Triggers, User Defined Functions (UDF), Views, Query Optimization, Database Partitioning, Index Management
- **Big Data & Analytics Platforms**: Hadoop, MapReduce, Snowflake, Snowflake Cortex Analyst, Apache Kafka, Redshift, 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, Go, Flask, RESTful APIs, OAuth 2.0, 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, Transformers, Natural Language Processing, Semantic Routing, Prompt Engineering
- **Visualization & Reporting**: Matplotlib, Seaborn, Power BI, Tableau, Custom Dashboards, KPI Tracking, Real-time Data Visualization, Automated Reporting Solutions, Business Metrics Tracking

WORK EXPERIENCE

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

- Architected and implemented HIPAA-compliant data pipelines for asthma patient research, leveraging transfer learning methodologies to engineer personalized ensemble prediction models achieving 88% accuracy in asthma onset forecasting - a 20% improvement over traditional classifiers and 12% enhancement compared to established neural network architectures.
- Secured comprehensive CITI Program certifications in research ethics and human subject protocols, establishing governance frameworks for sensitive medical data handling while authoring peer-reviewed research publications on novel machine learning approaches for healthcare analytics.
- Research Projects under review for publishing:
  - **Asthma Patient Research Project (South Korean Hospital Collaboration):**
    - Spearheaded advanced clustering analysis implementing diverse algorithms (K-means, DBSCAN, Affinity Propagation, BIRCH, Mean-Shift, OPTICS) to identify intricate patient cohort relationships while developing a comprehensive environmental analysis framework integrating weather and air quality data.
    - Implemented and optimized open-source Lag-Llama foundation model within a customized prediction framework to reconstruct missing values in time-series patient data, conducting rigorous comparative analysis against traditional imputation methods using multiple statistical metrics (MSE, RMSE, MAE, R^2, MAPE).
    - Performed systematic hyperparameter optimization for Lag-Llama context window configurations, effectively expanding usable training data segments by 37% while maximizing prediction accuracy on missing time-series values.
  - **Agricultural Computer Vision Project (Washington State Farmers Collaboration):**
    - Engineered sophisticated 3D visualization pipeline transforming 2D drone imagery into comprehensive volumetric models while implementing state-of-the-art YOLOv10 object detection architecture for automated plant counting with 89.8% accuracy.
    - Orchestrated comprehensive aerial data collection strategy utilizing drone technology and developed extensive training datasets through meticulous manual annotation of 2,000+ images with 7,000+ unique bounding boxes using open-source platforms (CVAT, Roboflow).
    - Conducted comparative performance analysis between YOLO (v8, v11, and v12) architectures, achieving superior 93.6% accuracy with YOLOv12 while integrating ML-Depth-Pro libraries for precise distance measurement and size estimation of detected fruits.

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

- Architected sophisticated ETL pipelines utilizing SSIS for multi-source data extraction and complex transformations, delivering 75% reduction in deployment costs and implementation time for Yardi's real estate business intelligence module.

- Engineered advanced BI visualization ecosystems with real-time data streaming and custom interactive dashboards, enhancing organizational decision-making while implementing rigorous data quality frameworks that reduced inconsistencies by 90%.
- Led cross-continental collaboration across 4 specialized teams spanning 3 time zones, managing enterprise-scale BI transformation initiatives for 50+ Middle Eastern clients while maintaining 99.99% system uptime.
- Designed extensible data ingestion architectures supporting multiple data types (historical, event-based, batch) from diverse sources, accelerating customer onboarding by 60% through automated migration workflows while enhancing processing performance by 40% through parallel computation strategies.
- Established robust monitoring infrastructure with proactive alerting mechanisms for critical ETL operations, resulting in 85% reduction in system downtime through early issue detection and resolution.
- Developed automated analytical reporting solutions for tracking KPIs and business metrics, generating approximately 20 hours weekly efficiency improvements by eliminating manual reporting processes.

Data Engineer   **Yardi Systems**, Pune, India   Nov 2016 - Mar 2019
- Engineered and deployed sophisticated ETL architectures leveraging Yardi's proprietary frameworks and SQL Server Integration Services to transform lease approval workflows, generating 50% enhancement in system utilization metrics and securing $3M+ in critical revenue retention.
- Architected high-performance data warehousing ecosystems incorporating dimensional modeling, fact table optimization, and advanced ETL processing to power real-time analytical dashboards monitoring 100+ key performance indicators across comprehensive property management domains.
- Mentored and technically guided 4 junior engineers while implementing robust exception handling frameworks and comprehensive logging mechanisms within ETL packages, ensuring data quality standards and maintaining exceptional system reliability metrics.
- Developed complex database objects including stored procedures, triggers, and optimized SQL queries while enhancing database performance through strategic index optimization, query tuning techniques, and sophisticated partitioning strategies for large-scale data environments.
- Designed and deployed incremental data processing methodologies for enterprise-scale datasets, achieving 15% reduction in overall processing latency while maintaining complete data integrity and orchestrating end-to-end SSIS package lifecycle management through deployment automation.
- Established comprehensive data governance standards with detailed documentation protocols for ETL processes and warehouse structural components, implementing automated testing infrastructure for data validation and continuous performance monitoring.
- Developed intuitive analytical interfaces and specialized reporting tools enabling stakeholders to effectively access and interpret complex property management datasets while instituting rigorous code review protocols and streamlined deployment processes.

Junior Data Engineer   **JJIT Fintech Pvt. Ltd.**, India   Feb 2015 - Oct 2016
- Architected and implemented sophisticated SSIS packages for complex ETL operations, managing the entire process from data extraction and transformation to package scheduling while developing versatile integration frameworks using SQL Server Data Tools (SSDT) to seamlessly process heterogeneous data sources including structured Flat Files, Excel spreadsheets, relational databases, and Raw File formats.
- Engineered comprehensive Data Mart solutions utilizing advanced data warehousing methodologies, establishing robust repositories for downstream analytical reporting while developing intuitive User Access Tools enabling stakeholders to generate customized ad-hoc reports and execute complex analytical queries within multidimensional Cube environments.
- Demonstrated expertise across complete Software Development Life Cycle (SDLC) including in-depth system analysis, architectural design, iterative development, comprehensive testing, and seamless implementation of large-scale data integration projects while implementing systematic diagnostic methodologies to identify and resolve complex technical challenges.
- Created sophisticated ETL architectures utilizing incremental load strategies to efficiently migrate data from staging environments to partitioned tables, leveraging advanced SQL Server partitioning methodologies for optimized performance with comprehensive error handling mechanisms and diverse logging methodologies to ensure data integrity.
- Developed complex database objects including Stored Procedures, Triggers, User-Defined Functions, optimized Indexes, normalized Tables, and specialized Views, employing advanced T-SQL techniques while creating specialized database views to surface critical information through user interfaces.
- Optimized query performance through systematic execution plan analysis and implementation of SQL Server query optimization techniques, significantly enhancing data load efficiency and overall system responsiveness while enforcing data consistency rules within Microsoft SQL Server environments.

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

- [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) *(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, Infra as code, NLP, QLoRA)*

- Architected and deployed a production-grade retrieval-augmented generation (RAG) system with semantic graph clustering techniques, attaining 90% BERT similarity scores and 89% precision/recall metrics while implementing sophisticated document chunking algorithms with overlapping segments to preserve contextual integrity across document boundaries.
- Engineered a high-performance vector embedding persistence layer utilizing Pinecone, maintaining 85% cosine similarity retention while optimizing dimensional reduction techniques for balancing query performance and semantic accuracy in high-volume retrieval.
- Designed and executed a comprehensive LLM evaluation framework facilitating systematic A/B testing across OpenAI (GPT-3.5), Anthropic (Claude 3.5), and Google (Gemini 1.5)models, implementing controlled prompting strategies with standardized performance metrics and developing quantitative evaluation pipelines for measuring hallucination rates, factual accuracy, and response relevance.
- Created instrumentation for tracking token utilization, latency profiles, and cost metrics across different model configurations, enabling data-driven decisions while designing a semantic routing architecture that analyzes query intent and complexity to dynamically select appropriate models, optimizing for both performance and computational efficiency.
- Fine-tuned Llama 2 (7B parameter model) utilizing QLoRA (Quantized Low-Rank Adaptation) techniques, achieving parameter-efficient domain adaptation while reducing computational requirements by 70% compared to full-parameter tuning and implementing gradient checkpointing and mixed precision training methodologies to optimize VRAM utilization.
- Constructed specialized training datasets through careful curation of domain-specific examples with robust validation processes, resulting in enhanced response quality for targeted enterprise use cases while contributing to the open-source AI community.
- Developed a distributed inference system with intelligent caching mechanisms that reduced average response latency by 65% while maintaining response quality, significantly enhancing user experience metrics and establishing a continuous deployment pipeline for iterative improvements.

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

- Architected and implemented a sophisticated containerized AI system with specialized agent architecture, engineering distinct analyzer and generator components that operate in concert to produce synthetic datasets with statistical fidelity to source data distributions while developing a modular, extensible framework that decouples data analysis from generation processes.
- Engineered robust error propagation mechanisms between system components with comprehensive logging and monitoring, implementing an advanced batch processing framework with dynamic batch sizing algorithms that automatically adjust processing parameters based on memory availability and CPU utilization patterns.
- Optimized data throughput with parallelized processing pipelines while maintaining strict data consistency guarantees, engineering memory-efficient data handling routines that minimize footprint during large dataset transformations through intelligent streaming and chunking methodologies.
- Leveraged Anthropic API and Claude 3.5 Sonnet through sophisticated prompt engineering techniques, developing specialized context-aware prompts that preserve statistical properties while implementing adaptive prompting strategies that dynamically adjust instruction specificity based on data complexity.
- Developed comprehensive validation pipelines to verify synthetic data distributions against source datasets, implementing statistical comparison methodologies to ensure generated data maintains essential characteristics of the original information.
- Engineered robust parameter validation, contextual help systems, and intelligent defaults while designing and implementing a secure API key management system with environment-based configuration and just-in-time authentication.
- Published the containerized solution to Docker Hub with comprehensive documentation, continuous integration testing, and versioned releases, facilitating widespread adoption while maintaining quality control over distributed components.

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

- Architected and implemented a sophisticated cloud-native SaaS platform for enterprise workforce management, leveraging AWS services and microservices architecture while designing specialized data partitioning strategies within DynamoDB that optimized for common access patterns while minimizing read/write capacity consumption, resulting in 40% lower operational costs.

- Engineered multi-tiered storage architecture that intelligently routes data between hot and cold storage tiers based on access frequency analysis, implementing automated lifecycle policies that reduced storage costs while maintaining sub-second query performance without compromising data availability.
- Developed a comprehensive RESTful API ecosystem with granular endpoint permissions and rate-limiting mechanisms, implementing OAuth 2.0 authentication with custom scope definitions while engineering a robust API versioning system with backward compatibility guarantees to protect client implementations.
- Implemented containerized deployment architecture using Docker with optimized multi-stage builds, reducing image sizes by 65% while configuring advanced Kubernetes deployments on AWS EKS with automated horizontal pod scaling based on custom metrics and sophisticated readiness and liveness probes.
- Designed event-driven processing workflows leveraging AWS Lambda functions with carefully calibrated memory allocations and execution timeouts, implementing concurrency management strategies that delivered 5x throughput improvements for batch operations.
- Engineered a fault-tolerant dead-letter queue system with automatic retry mechanisms and intelligent backoff strategies, ensuring reliable data processing even during intermittent service disruptions while enabling secure third-party integrations without service degradation.

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

- Architected a sophisticated marketing campaign optimization system using ensemble machine learning techniques (Decision Tree, KNN, Random Forest), achieving 86% prediction accuracy in identifying high-conversion customer segments while developing custom evaluation metrics aligned with business objectives.
- Engineered comprehensive model interpretation methodologies that translated complex algorithmic patterns into actionable marketing insights and strategic recommendations for targeted campaign optimization.
- Implemented advanced class imbalance mitigation strategies combining SMOTE and Random Under Sampling techniques, significantly improving minority class prediction without compromising overall model accuracy.
- Designed robust feature engineering pipelines that transformed raw customer interaction data into predictive indicators, creating compound features that captured complex behavioral patterns while systematically eliminating noise variables.
- Developed sophisticated exploratory data analysis workflows with custom visualization frameworks using Matplotlib and Seaborn, implementing multi-dimensional analysis methodologies that revealed previously undetected patterns across diverse market segments.
- Engineered automated data quality systems incorporating statistical outlier identification algorithms and anomaly detection, substantially improving pipeline integrity prior to model training phases.
- Implemented comprehensive model optimization protocols combining stratified cross-validation with hyperparameter tuning systems using GridSearchCV and custom scoring functions, establishing automated workflows that systematically identified optimal configurations for production deployment.

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

- Architected and developed an advanced SQL query generation system leveraging Snowflake Cortex Analyst, creating a sophisticated conversational interface that transforms natural language questions into optimized SQL queries for non-technical users.
- Engineered a comprehensive semantic model framework with YAML configurations defining logical tables, dimensions, measures, and custom expressions, enabling high-precision query generation through detailed metadata mapping of complex database.
- Designed and deployed a production-grade Streamlit chatbot interface with robust error handling, request management, and sophisticated session state capabilities, creating an intuitive experience for natural language query composition while maintaining secure connections throughout user sessions.
- Implemented advanced data ingestion workflows for multiple revenue datasets with optimized ETL processes, precise data type handling, and error management protocols to ensure data integrity across the analytical ecosystem.
- Engineered sophisticated API integration with Snowflake Cortex Analyst REST endpoints, implementing secure token-based authentication and efficient request payload structuring while developing intelligent caching mechanisms that significantly reduced API call frequency.
- Designed a verified query repository system for capturing validated SQL patterns, enhancing query accuracy and performance by leveraging proven structures for similar natural language inputs while providing comprehensive error handling with detailed logging.

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.