

Yash Rupani

 +1 541-250-1204 |  rupaniyash1818@gmail.com |  [LinkedIn](#) |  [GitHub](#) |  [Portfolio](#) |  United States

SUMMARY

Data Engineer with a Master's in Computer Science and **2+ years of hands-on experience building scalable, cloud-native data platforms** across AWS and GCP. Proven ability to **streamline ETL pipelines, minimize infrastructure costs, and accelerate data processing at scale**. Strong background in **Apache Spark, Python, distributed systems, and AI-driven data architectures**, with real-world experience supporting **Agentic AI, RAG systems, and large-scale analytics workloads**.

EXPERIENCE

Research Assistant – Data Engineering Focus <i>Oregon State University</i>	Oct. 2025 – Dec. 2025 <i>Corvallis, OR</i>
<ul style="list-style-type: none">Engineered an end-to-end autonomous ETL pipeline for unstructured multimedia data, leveraging Python-based preprocessing and vectorization to eliminate 40% of manual data preparation for Agentic AI training workflows.Refined backend data structures and indexing strategies for AI Agents, slashing retrieval latency by 25% and boosting real-time inference performance.Developed and enforced robust data validation and sanitization frameworks to ensure 100% data integrity while enabling scalable ingestion of terabyte-scale datasets.Tech Stack: Python, ETL Pipelines, Data Validation, Indexing Algorithms, AI Data Engineering, Distributed Systems.	
AI Agent Graph RAG / Machine Learning Scientist Intern <i>GrantAide</i>	Jul. 2025 – Sep. 2025 <i>Remote</i>
<ul style="list-style-type: none">Architected and streamlined scalable backend systems using Flask and Google Firestore, restructuring database access patterns to curtail API latency by 30% during peak usage.Implemented FAISS-based vector search pipelines to enable semantic retrieval, increasing search relevance by 40% compared to traditional keyword matching.Orchestrated multi-cloud deployments across AWS and GCP, achieving 99.9% uptime while reducing cross-environment latency by 25%.Tech Stack: Flask, FAISS, Vector Databases, RAG Systems, AWS, GCP, Cloud Architecture.	
Teaching Assistant – Integrated Business Analytics <i>Oregon State University</i>	Apr. 2025 – Jun. 2025 <i>Corvallis, OR</i>
<ul style="list-style-type: none">Directed and supervised Python-based data analytics pipelines for industry-sponsored projects, ensuring 100% on-time delivery for enterprise clients including Port of Portland.Designed and scripted data quality validation checks using Pandas and NumPy, bolstering dataset reliability by 25% for downstream modeling.Mentored 15+ students on requirement gathering, stakeholder communication, and technical storytelling, leading to higher client satisfaction scores.Tech Stack: Python, Pandas, NumPy, Data Analytics, Client Communication, Mentorship.	
Senior Systems Engineer – Data Engineering Support <i>Infosys</i>	Jun. 2021 – Jun. 2023 <i>Pune, India</i>
<ul style="list-style-type: none">Spearheaded modernization of enterprise data quality pipelines, eliminating 10+ hours/week of manual data intervention.Designed and deployed high-performance ETL workflows to process large-scale system monitoring data, compressing processing runtime by 40% per cycle.Constructed dynamic dashboards and monitoring solutions from semi-structured logs, replacing manual log analysis with scheduled workflows.Tech Stack: ETL Design, Apache Spark, Data Quality Engineering, Automation, Monitoring Systems.	

PROJECTS

Crypto Sentinel: Real-Time Market Monitor Live Demo Spark, Redpanda, DuckDB, Streamlit
<ul style="list-style-type: none">Engineered a streaming pipeline processing Coinbase WebSocket data via Redpanda and Spark, solving the small file problem with DuckDB.Optimized dashboard performance using Streamlit Fragments, eliminating UI jitter during high-frequency (1Hz) updates.Built a robust Matrix-style freeze feature, allowing users to pause live streams for deep-dive analysis without breaking ingestion.
Shop Pulse: Real-Time E-Commerce Lakehouse Live Demo Spark, Kafka, Python, Docker
<ul style="list-style-type: none">Architected a Lakehouse pipeline using Kafka and Spark Streaming, achieving sub-second latency for live visualization.Implemented a fault-tolerant storage layer using partitioned Parquet files, optimizing query performance for downstream analytics.Containerized the producer-consumer architecture via Docker to simulate high-volume traffic without resource contention.

TECHNICAL SKILLS

Programming: Python, SQL, C++, Bash

Data Engineering: Apache Spark, PySpark, Airflow, Kafka, Snowflake, Hadoop, Hive, Presto

Cloud Platforms: AWS (S3, Lambda, Glue, EMR, Redshift), GCP (BigQuery, Firestore)

Databases: PostgreSQL, MySQL

ML & AI Systems: FAISS, Vector Databases, RAG Architectures, NLP Pipelines

Tools: Docker, Git, Tableau

EDUCATION

Oregon State University

Master of Engineering in Computer Science

Corvallis, OR

Sep. 2023 – Dec. 2025

Pandit Deendayal Energy University

Bachelor of Technology in Electrical Engineering

Gujarat, India

Aug. 2017 – Jun. 2021