

Yash Rupani

☎ +1 541-250-1204 | ✉ rupaniyash1818@gmail.com | [in 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

Oct. 2025 – Dec. 2025

Oregon State University

Corvallis, OR

- **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

Jul. 2025 – Sep. 2025

GrantAide

Remote

- **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

Apr. 2025 – Jun. 2025

Oregon State University

Corvallis, OR

- **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

Jun. 2021 – Jun. 2023

Infosys

Pune, India

- **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

🔗 [YouTube Trending Video Data Engineering Pipeline](#) | *AWS S3, Lambda, Glue, Athena, PySpark*

- **Architected** a serverless AWS data lake ingesting datasets representing **102B+ video views**, converting raw JSON into optimized Parquet formats to **cut storage costs by 65%**.
- **Developed and fine-tuned** PySpark ETL jobs on AWS Glue, improving Athena query efficiency and **driving down scan costs by 50%**.
- **Delivered** interactive Amazon QuickSight dashboards **unlocking** cross-regional insights across **40,000+ trending videos**.

🔗 [Financial Markets News Sentiment Analysis \(NLP\)](#) | *Python, NLP, FinBERT, NLTK*

- **Built a continuous** NLP pipeline ingesting **500+ real-time financial news articles per run** for market sentiment analysis.
- **Integrated** FinBERT transformer models to achieve **90% sentiment classification accuracy**, outperforming traditional lexicon-based methods.
- **Translated** unstructured text into actionable insights via data visualization, **facilitating** clearer interpretation of market volatility trends.

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 <i>Master of Engineering in Computer Science</i>	Corvallis, OR <i>Sep. 2023 – Dec. 2025</i>
Pandit Deendayal Energy University <i>Bachelor of Technology in Electrical Engineering</i>	Gujarat, India <i>Aug. 2017 – Jun. 2021</i>