SANCHIT VIJAY

sanchit.aiwork@gmail.com | 202-391-3369 | linkedin.com/in/sanchit-vijay | github.com/sanchitvj

WORK EXPERIENCE

Opal HTM, Washington, DC — **Data Engineer** Oct 2023 – Present

- Architected big data pipeline for medical device analytics following **lakehouse** architecture, reducing predictive analysis latency by **85%**.
- Engineered serverless data ingestion system using AWS **Lambda** and **ECS Fargate**, processing **50M+** data points per trial with **99.9%** data quality through **DynamoDB** tracking.
- Optimized **Spark**-based **ETL** pipelines on **AWS EMR**, reducing processing time by **80%** and enabling seamless integration with **Redshift** for analytics workloads.
- Implemented data quality monitoring using **Airflow**, AWS **Glue** and **Iceberg** table format, enabling **40%** faster query performance through **Athena** for ad-hoc analysis.
- Established infrastructure automation using GitHub Actions and **Terraform**, achieving **3x** deployment frequency through containerized microservices (**ECR**), reducing cloud costs by **30%**.

Bytelearn, India — **Data Engineer** July 2021 – July 2022

- Built **AWS Glue** workflows to automate image metadata extraction, improving dataset accuracy by **25%** for downstream analytics.
- Designed annotation tool backend with **FastAPI** and UI with **Streamlit**, reducing manual work by **80%** for image data and improving rendering time by **60%** for video content.
- Developed algorithms for image data generation, ingestion of unstructured data, cutting development time by **70%** through modularization.
- Employed Docker-based deployment environment with **Agile workflows**, ensuring reproducibility across **5+** systems and accelerating cross-functional collaboration.

TECHNICAL PROJECTS

Real-time Sports Betting Analytics Engine (BetFlow) Nov 2024 – Jan2025

- Architected sports betting platform using **Lambda** architecture, processing **400K+** records daily with **Kafka**, **Spark** Streaming, and **Druid** on local infrastructure enabling sub-second market analysis.
- Engineered data pipelines integrating real-time streams (games, odds, weather) with **OLAP**-based historical analysis using **Snowflake** and **DBT**, reducing analytics latency to **5 seconds**.
- Accelerated analytics using incremental strategy, **SCD** Type-2, and **CDC** patterns in DBT, reducing daily warehouse compute cost by **60%** while enabling betting market inefficiency detection.
- Orchestrated batch ETL using **Airflow** and optimized Snowflake external tables with **Glue** catalog integration, reducing warehouse storage costs by **90%** while maintaining query performance for **1TB+** data.
- Designed multi-sport **Grafana** dashboards handling **1M+** daily events across betting analytics and market trends, enabling stakeholders to analyze patterns with sub-**5 second** refresh rate.

Good Retrieval Augment Generation (GRAG) Jan 2024 – May 2024

- Developed GRAG, zero-cost Retrieval-Augmented Generation (**RAG**) package enabling high-accuracy searches across diverse document formats, prioritizing data privacy.
- Created user-friendly **5-line** code framework for RAG, simplifying local or HuggingFace LLM integration, enabling end-to-end functionality for custom databases and vector stores.
- Automated build, coverage, and testing processes using **CI/CD** (Jenkins and GitHub Actions) to streamline deployment and ensure long-term maintainability of GRAG.

TECHNICAL SKILLS

- Programming & Data Analysis: Python, R, SQL, NoSQL, Postgres, Tableau, MS Excel, PySpark
- Machine Learning: TensorFlow/Keras, PyTorch, Langchain, MLflow, FastAPI, Streamlit, REST APIs
- Data Engineering: Snowflake, Databricks, Apache (Spark, Kafka, Flink, Airflow, Druid, Superset, Hive, Hadoop), DBT, Polars, Dagster, Grafana, Trino, DuckDB
- Cloud Services (AWS): EMR, Glue, Athena, Redshift, IAM, S3, EC2, ECS, ECR, Lambda, Sagemaker, DynamoDB
- DevOps & CI/CD & Agile: GitHub Actions, CircleCI, Terraform, Jenkins, Docker, Jira, Confluence

EDUCATION

- **The George Washington University**, Washington, D.C.
  Master of Science in Data Science; GPA: 3.87