RAHUL BAID

rahulbaid0110@gmail.com | rahulbaid.co | github.com/rahulb99 | linkedin.com/in/rahul-baid

EDUCATION

Texas A&M University, Masters in Computer Science (GPA: 4.0/4.0)   Aug 2024 - Current
- **Courses** - Information Retrieval, Data Visualization, Reinforcement Learning, Parallel Computing
- **Research** - Reinforcement Learning/Trustworthy LLMs at FLAIR Lab
- **Teaching Assistant** - Natural Language Processing (Spring '25)

National University of Singapore, Bachelor of Computing   Aug 2017 - May 2021
Computer Science (Major) with a specialization in Artificial Intelligence, **Statistics** (Minor)
- **Teaching Assistant** – Deep Learning, Natural Language Processing, Computer Organization
- **Internships** – Software Engineer [HP, Rakuten, Treatsure], Machine Learning Engineer [Dathena]
- **Research** - Financial NLP (data extraction, knowledge graph, QA with FinBERT) at NUS FinTech Lab

Georgia Institute of Technology, Student Exchange Program   Jan 2020 - May 2020

EXPERIENCE

PropertyGuru (Singapore), Data Engineer, Data Platform Team   Jul 2021 - Aug 2024
- Architected a high-throughput **geospatial data processing pipeline** 10M+ daily location pins for a Mall Analytics SaaS product, enabling real-time foot traffic insights. Tech: AWS EMR, Redshift, Glue, Lambda, OpenSearch, Iceberg.
- Created the **Address Intelligence System** by unifying fragmented datasets, building data models and applying necessary transformations, into a single source of truth, accelerating query performance by **20%** and reducing onboarding time for analysts by 30%. Tech: BigQuery, dbt, FastAPI, Elasticsearch, GraphQL, OpenStreetMap.
- Engineered a **reverse ETL pipeline framework** achieving sub-5-minute operational data syncs, **saving $30K/year** by replacing third-party SaaS. Tech: Kafka, dlt, Airflow, Polars, BigQuery, DuckDB.
- **Automated 85%** of data validation tasks by building a data quality framework, integrated with an AI copilot, leading to **30% fewer errors**. Tech: dbt, Great Expectations, SodaSQL, OpenAI, SQLGlot.
- **Slashed BigQuery costs by 30%** annually through storage and query-optimization strategies, while introducing DuckDB for lightweight transformations.

PROJECTS

- **theLook E-commerce Analytics** - Streaming and batch ETL pipelines to ingest, process and analyze high-volume and velocity data. Tech: AWS (Lambda, ECR, RDS), Confluent (Kafka, Flink), Databricks (Spark, Delta Lake). https://github.com/rahulb99/ecommerce-analytics
- **Real-time analytics of transit feeds** - Streaming data pipeline using Kafka and Flink to analyze Austin buses' trip delays, overcrowding, sub-optimal routes and schedules, and road accidents. https://rahulbaid.co/posts/project-transit-feeds/
- **RAID** - Web app to visualize historical and near real-time insights into traffic accidents in Austin. Tech: D3.js, Mapbox, React, Postgres, Heroku. https://github.com/rahulb99/dataviz-raid
- **Analysis of small, local LLM safety** - Empirical evaluation of jailbreaking attacks and defenses against LLMs deployed at the edge. https://github.com/rahulb99/trustworthy-nlp

SKILLS

Programming	Python, SQL, Java, JavaScript, PyTorch
Cloud	AWS, GCP, Databricks, Confluent, Docker/Kubernetes, Terraform
Tools	Airflow, Kafka, Flink, Spark, Trino, Grafana, dbt, Datahub, Github Actions
Databases	Elasticsearch, MongoDB, PostgreSQL, Clickhouse, Pinot, Redis
AI	LangGraph, LangChain, Agno, Opik, Langfuse