

BHAVYA UPADHYAY

Irvine, CA | +1 (949) 617-7184 | officiallybhavya@gmail.com
linkedin.com/in/bhavyaupadhyay | github.com/bhavyaupadhyay | bhavyaupadhyay.site

SUMMARY

Master of Data Science candidate at UC Irvine (full merit scholarship) with one year of production data engineering experience at TCS, where I shipped AWS ETL pipelines processing 10M+ records monthly. I build end-to-end data and ML systems: cloud ETL, multi-agent LLM systems and RAG applications, and deep-learning models served behind APIs. Strongest in Python, SQL, and AWS, with a focus on turning messy data into reliable, monitored pipelines.

EDUCATION

University of California, Irvine

Master of Data Science | Full Merit Scholarship (fee remission + stipend)

Irvine, CA

Expected Dec 2026

SRM Institute of Science and Technology

B.Tech, Electronics & Communication Engineering (Data Science specialization)

Chennai, India

Sept 2020 - May 2024

TECHNICAL SKILLS

Languages: Python, SQL, R, Bash

Data Engineering & Cloud: AWS (Lambda, Glue, Athena, S3, CloudWatch), Airflow, dbt, Snowflake, Kafka, ETL pipeline design, Docker, FastAPI, Streamlit, Git, Terraform, Kubernetes

ML & AI: PyTorch, scikit-learn, XGBoost, LightGBM, Hugging Face, model evaluation, ablation studies, hyperparameter tuning

NLP & LLMs: LangChain, RAG pipelines, Sentence Transformers, prompt engineering, semantic search, vector databases (Milvus)

Data & Analytics: Pandas, NumPy, PostgreSQL, Supabase, Tableau, Power BI, Matplotlib, A/B testing, statistical modeling, cohort analysis

EXPERIENCE

Data Analyst | University of California, Irvine (Graduate Division)

Sep 2025 - Present

- Automated data integration across 5+ institutional systems (enrollment, funding, admissions) with Python and SQL ETL, removing ~40 hours/month of manual reconciliation and resolving duplicate and mismatched records.
- Built retention-risk and graduation-forecasting models across 20+ graduate programs to inform funding and resource-planning decisions.
- Designed 8+ Tableau and Python KPI dashboards for cohort analysis and retention tracking, cutting reporting turnaround by ~30%.

Data Engineer (promoted from Assistant Systems Engineer) | Tata Consultancy Services (Air Canada)

Jun 2024 - Jul 2025

- Engineered Python ETL pipelines on AWS Lambda processing 10M+ flight and booking records monthly, cutting data latency ~50% and enabling near-real-time operational analytics.
- Optimized 50+ Athena SQL queries behind enterprise reporting, reducing average query runtime ~40% and speeding data delivery to 10+ downstream teams.
- Built 12+ Grafana dashboards on AWS CloudWatch with automated anomaly alerting, cutting incident triage time ~35% and improving pipeline SLA compliance to 99.9%.

PROJECTS

EDGAR-X | Multi-Agent Financial Intelligence System | Python, Snowflake, dbt, Airflow, XGBoost, Anthropic API, Terraform, Kubernetes

2026

- Built a 7-layer pipeline ingesting SEC EDGAR filings and FRED macro data for 498 S&P 500 companies into Snowflake (7,863 multi-year XBRL fundamentals rows, 4,758 parsed 10-Ks) via key-pair auth and an idempotent, checkpointed backfill; 75 dbt tests surfaced four real data-extraction bugs at scale
- Trained an XGBoost revenue-direction model as a ranked screen (ROC-AUC 0.726 vs 0.500 baseline) with SHAP and a backtested calibration loop; built a multi-agent LLM tier producing source-grounded research memos with code-supplied provenance and LLM-as-judge evaluation
- Shipped a live public Streamlit dashboard on a static snapshot (free, permanent) and deployment-ready Terraform/Kubernetes IaC with GitHub Actions CI; live demo: edgar-x-26rm39zy6c8fjzyb9pxdt.streamlit.app

Flightline | End-to-End Flight Data Pipeline | Python, Airflow, Snowflake, dbt, Streamlit, Docker, GitHub Actions

2026

- Built a batch ingestion pipeline loading US BTS On-Time Performance data into Snowflake via Airflow DAGs; idempotent partition-delete-before-COPY loads with MATCH_BY_COLUMN_NAME prevent drift and duplication on reruns
- Modeled raw data into staging, fact, dimension, and aggregate layers with dbt; 26/26 tests passing including two custom business-rule tests; CI lints SQL with sqlfluff and runs dbt build against an ephemeral schema on every push
- Architected OpenSky live streaming path with OAuth2 token refresh and Kafka decoupling for operational flight positions; served carrier on-time rates and worst-route delay analytics on a live Streamlit dashboard

Signal Miner | LLM-Powered Market Intelligence | Python, LangChain, Supabase, Docker, Streamlit

2025

- Engineered an LLM pipeline (LangChain prompt chaining + RAG) that turns unstructured market data from 250+ sources into clean, structured records
- Modeled a Supabase store with deduplication and vector retrieval and containerized the full stack with Docker for reproducible deployment
- Built a Streamlit interface to query and visualize extracted signals, with prompt versioning and output validation to catch malformed LLM responses before they reached the store