

AAKASH MADABHUSHI

+1 408-210-6122 | aakash.vardhan15@gmail.com | San Francisco Bay Area | Available Dec. 2026
linkedin.com/in/aakash-varadhan | github.com/aakashvardhan | aakashvardhan.github.io

SUMMARY

Build the pipelines that feed dashboards and models, and the checks that keep bad data out of both: 5M+ rows a day through multi-tenant ETL in Databricks and PySpark, a 43% cut in p95 dashboard query latency, and 27.1M network flows condensed into a graph a model could train on. Batch and streaming, with schema, freshness, and integrity gates in front of everything.

EXPERIENCE

Graduate Research Assistant

Mar. 2026 – Present

SJSU Applied Data Science Lab

San Jose, CA

- Published **49,633 frames across 50 episodes** as an open dataset by building a LeRobot v3 collection pipeline that synchronizes 6-DoF joint states and action trajectories with dual 640x480 RGB streams at 30 FPS.
- Cleared every published episode with **zero integrity failures** by developing pre-training quality gates over Parquet metadata and H.264 video that validate episode and frame alignment, video decodability and resolution, and NaN-free action and proprioceptive tensors.
- Let operators run a 100-trial evaluation protocol unattended by building a resumable harness with automatic return-to-start, stale-cache guards, video-measured placement error, and per-trial success and failure logging.

Software Engineer

Jun. 2023 – Jun. 2024

Sapaad

Dubai, UAE

- Served tenant-facing analytics on **5M+ rows per day** by shipping production Python ETL/ELT pipelines in Databricks and PySpark that extracted recommendation-model features across a multi-tenant SaaS platform.
- Cut p95 dashboard query latency **43%** (12.4 s to 7.1 s) across 10+ Spark jobs serving **75+ tenant dashboards** by profiling execution plans, rewriting SQL window functions, and placing targeted caches.
- Caught 14 schema, freshness, and ingestion anomalies automatically by building a Delta Lake data-quality framework with schema validation and freshness checks across every ingestion pipeline.

Research Assistant

Sep. 2019 – Mar. 2020

Illinois Institute of Technology

Chicago, IL

- Identified water-related disaster events from live social data by automating collection through the Tweepy API and generating FastText embeddings over the harvested corpus.
- Reached 85% classification accuracy, a further 7% over the untuned baseline, by training an SVM over those embeddings and tuning it with GridSearchCV.

PROJECTS

NYC Taxi Data Platform: Batch and Streaming | *Airflow, PySpark, dbt, Snowflake, Kafka, Docker* Nov. – Dec. 2025

- Served the same trip data to both dashboards and live monitoring by building it twice: a scheduled Airflow, dbt, and Snowflake warehouse joined against weather, and a Kafka streaming path over identical sources.
- Published rolling five-minute metrics on the streaming path by validating each trip on arrival and routing rejects to a dead-letter topic rather than dropping them silently.
- Modeled the warehouse in dbt so dashboard-ready tables are rebuilt from raw trips on every scheduled run rather than patched in place.

Network Threat Detection Pipeline | *Airflow, dbt, Snowflake, PyTorch Geometric, Pinecone* Feb. – May 2026

- Condensed **27.1M raw Zeek network flows** into a host-pair connection graph a model could train on, by building an Airflow, dbt, and Snowflake medallion pipeline over the raw log volume.
- Supported **0.90 F1** link classification (0.94 precision, 0.87 recall) on 329 held-out host pairs by engineering the aggregated features the graph neural network consumes.
- Delivered retrieval-grounded alert explanations at **0.87 s** median end-to-end latency by serving graph context from a Pinecone vector index into a language model step.

U.S. Flight Delays: Batch Analytics to Real-Time Prediction | *PySpark, Spark SQL, Kafka, Streamlit* Dec. 2025

- Processed **9.5M flight records** from 15 monthly BTS files totalling 3.1 GB into a single analysis-ready table with PySpark and Spark SQL.
- Moved the batch-trained model onto a Kafka Structured Streaming path so each arriving flight receives a delay probability within seconds of ingestion.

- Surfaced the result through a Streamlit dashboard reading from the streaming sink, so batch analysis and live prediction present from one interface.

Stock Market Analytics Pipeline | *Airflow, dbt, Snowflake, Docker*

Oct. – Nov. 2025

- Kept a full revision history of every price the warehouse ingests by modeling the load as dbt snapshots, so a restated figure stays recoverable rather than being overwritten in place.
- Loaded daily stock prices into Snowflake on an Airflow schedule and computed the indicators traders use as dbt models downstream of the raw layer.
- Packaged scheduler, warehouse connection, and transformations in Docker so the pipeline comes up identically on any machine.

TECHNICAL SKILLS

Pipelines & Orchestration: Apache Airflow, dbt, Apache Kafka, Spark Structured Streaming, batch and streaming ETL/ELT, medallion architecture, dead-letter queues, backfills, scheduling

Processing & Warehousing: Apache Spark, PySpark, Spark SQL, Databricks, Snowflake, Delta Lake, Parquet, SQL, MySQL, MongoDB, window functions, query and execution-plan tuning, caching

Data Quality & Modeling: schema validation, freshness checks, anomaly detection, integrity gates, dimensional and medallion modeling, dbt tests and snapshots, slowly changing data

Languages: Python, SQL, Bash, TypeScript

Infrastructure: Docker, Docker Compose, Git, Linux, GitHub Actions, AWS EC2, FastAPI, Redis, Pinecone

EDUCATION

San Jose State University

M.S. in Applied Data Science

Expected Dec. 2026

San Jose, CA

Illinois Institute of Technology

B.S. in Artificial Intelligence

May 2023

Chicago, IL