Rohan Anand

anandro@bu.edu | rh-anand.vercel.app | linkedin.com/in/rohan-h-anand | github.com/rohan-anand21 | (603) 233-4670

Experience

Data Engineer Mar 2025 – Present

Dataeconomy *Charlotte, NC*

- Built an implicit-feedback ALS recommender for a data marketplace; improved top-10 precision by ∼50% vs a popularity baseline via Optuna hyperparameter tuning.
- Productionized model as a Dockerized FastAPI service on AWS ECS and exposed a REST endpoint to serve predictions.
- Created a Jenkins CI/CD pipeline to automate retraining, validation, and deployment of the recommender service.
- Implemented Kafka → Snowpipe → Snowflake near-real-time ingest for low-latency dataset updates feeding downstream analytics and recommendations.
- Designed data quality profiling in Snowflake/Snowpark; computed column/row metrics, auto-segregated valid vs invalid records, and generated detailed data-quality reports for stakeholders.
- Rewrote an internal data-validation package using Great Expectations in a PySpark ETL pipeline; added new checks and constraint categories and reduced runtime ∼70% on large datasets.
- Staged and loaded Snowflake-managed Iceberg tables from S3 and implemented SCD Types 0–4 using SQL + dbt for complete historical data management.
- Building an internal FastAPI app for no-code Airflow orchestration with integrations to Talend, Informatica, Confluence, and Teams; used GraphQL for front-end ingestion and the platform's REST API for backend execution.

AI Engineer Sep 2024 – Jan 2025

BU Spark! *Boston, MA*

- Ingested & normalized officer records (2011–2024) from heterogeneous spreadsheets (∼ 14 files, ∼ 50k rows, ∼ 30 columns). Built a repeatable pipeline to clean, dedupe, and entity-resolve officers/cases; produced tables and incremental update routines for new data drops.
- Implemented an embeddings pipeline with OpenAI and a Chroma vector index (∼ 1000 vectors, dim = 1,536). Applied chunking (∼ 500 tokens/chunk) and metadata enrichment (officer_id, allegation type, year) to improve retrieval quality and maintain traceability.
- Delivered a FastAPI service exposing /search and /recommend endpoints: Chroma → retrieval → OpenAI generation, orchestrated with LangChain; supports CSV uploads for new records and writes request/response audit logs for review.
- Client reported ∼ 80% reduction in initial investigation time after adopting the workflow; offline evaluation improved ∼ 40% on Recall@10/MRR via feature augmentation (e.g., prior-offense context) and targeted prompt templates.
- Built geospatial maps in Python to help target which communities to review allegations of police officers (∼ 23 areas, ∼ 13 time windows); overlaid counts/rates and basic filters (e.g., year, allegation type) to guide client review sessions.

Data Services Intern May 2024 – Aug 2024

Axis Technology, LLC *Boston, MA*

- Built a scalable synthetic-data generator in Python (1k → 100k+ rows) for PII detection; substantially improved offline performance on a held-out real dataset; maintained stable runtimes.
- Wrote a schema parser/labeler to read JSON database schemas and tag column semantic types for downstream features.
- Authored tests to evaluate OpenSearch query relevance for table-similarity search; documented metrics and cases.

Data Analyst Intern Jun 2022 – Aug 2022

AS Insurance Agency *Manchester, NH*

- Consolidated customer and policy data (2.5k+ docs) with Pandas; wrote Snowflake SQL to segment customers; built 10–15 Power BI/Tableau dashboards for renewal targeting.
- Helped streamline renewal outreach process; contributed to 95% client retention over the period.

Projects

Analyzing Boston's 311 Service Requests Sep 2023 – Dec 2023

- Built an automated daily API ingestion for Boston 311 (2.7M+ records across 12 years); normalized raw responses into analysis-ready tables and added basic deduping and schema validation to keep the dataset clean.
- Developed interactive analyses in Jupyter (ipywidgets) and a Power BI map layered with the CDC Social Vulnerability Index (SVI); examined trends by neighborhood, request type, submission source, and resolution time.
- Produced and presented a Power BI report showing differences across income levels; documented the data structure and the daily refresh schedule.

Skills

Languages: Python, SQL; R (basic) | **Data Engineering:** Airflow, Kafka, PySpark, dbt, Great Expectations, FastAPI, Data Modeling (SCD 0–4, Iceberg) | **Cloud & Warehousing:** AWS (ECS/ECR, S3), Snowflake (Snowpipe, Snowpark, Iceberg); Azure (basic) | **Databases & Search:** PostgreSQL, OpenSearch | **ML/RAG:** Pandas, NumPy, scikit-learn, Optuna, PyTorch, OpenAI API, LangChain, Chroma | **Containers & CI/CD:** Docker, Jenkins, Git | **BI & Tools:** Power BI, Tableau, Postman, Confluence, Jira

Certifications

AWS Certified Cloud Practitioner *Issued May 2025*

Education

Boston University Boston, MA

B.S. in Data Science | Dean's List: 2022 - 2025 | DS Undergraduate Tutor (Mar 2023 - Dec 2024) *09/2021 – 01/2025*