

ANANYA CHEMBAI

(412) 450-7128 · ananyachembai@gmail.com · [linkedin.com/in/ananyachembai](https://www.linkedin.com/in/ananyachembai)
github.com/ananyachembai · www.ananyachembai.com · U.S. Citizen

PROFESSIONAL EXPERIENCE

PNC Bank

Data Engineer

Data Engineer Associate

Pittsburgh, PA

Feb 2026 - Present

May 2025 - Feb 2026

- Promoted within **nine months** after receiving an Exceeds Expectations performance rating; **one of three engineers promoted from a 28-person team**.
- Owned the two-month resolution of a monthly-statement delivery failure caused by lengthy Oracle materialized-view refreshes, from solution design through production release. Redesigned the entire data architecture to use daily snapshot tables for each view, cutting refresh time by **~75%**.
- Engineered an **agentic workflow** to automate pre- and post-refresh validation for an **800GB** broker-dealer trading database, reducing hands-on validation time by **~88% per monthly refresh**.
- Delivered a Canadian-entity reporting system for two new AML sanctions-screening requirements, **translating regulatory compliance needs into SQL logic**, escalation-ready reporting, and production workflows for downstream review.
- Owned delivery of a T+1 allocations report, **working directly with clients** to translate reporting needs into a single output for allocated and unallocated trades. The solution **eliminated manual reconciliation** and enabled timely reporting under T+1 settlement deadlines.

UBS

Financial Analyst Intern

Hyderabad, India

Apr 2024 - Jul 2024

- Migrated month-to-date P&L reporting from legacy SQL to **PySpark on Databricks**, accelerating query execution and reducing overall report-generation time by **~67%**.
- Modernized legacy dataset delivery by moving releases to **Azure** and implementing **Git-based CI/CD pipelines**, creating a repeatable, version-controlled deployment process.

PROJECTS

Predictive Optimization of Reddit BigQuery Workloads

CMU Capstone | GCP

- Partnered directly with Reddit to analyze **4.1M BigQuery jobs** using PySpark, uncovering zero-output workloads that consumed **13.8% of total slot time** and multi-input jobs with **over 90% lower efficiency**.
- Built **regression models** to prioritize workload-optimization opportunities and translated the findings into recommendations for batching metadata jobs, rescheduling heavy workloads, and right-sizing slot reservations. Presented the recommendations **directly to stakeholders**.

MatchMadeInMed: Rare-Disease Clinical-Trial Prototype

Harvard Rare Disease Hackathon

- Built a rare-disease clinical-trial prototype that matched patient profiles to trial eligibility criteria. Used **conditional GANs** to generate **synthetic control groups** and **XGBoost** to predict placebo and treatment responses.
- Achieved **R² 0.96 in predicting patient response on held-out data**. Delivered an interactive Streamlit workflow for reviewing patient matches and comparing simulated placebo vs. treatment responses.

Retrieval-Augmented Research Assistant

CMU | NLP & LLM Evaluation

- Built a **RAG system** over academic papers using **SentenceTransformers** and **ChromaDB**. Created a question-based semantic index to retrieve evidence for grounded GPT-4o-mini responses.
- Added an **"IDK" abstention rule** for questions unsupported by the retrieved context, limiting responses to available evidence instead of generating ungrounded answers.
- Evaluated generated answers against reference responses using **BERTScore**. Achieved **0.84 precision, 0.92 recall, and 0.87 F1** on held-out test queries.

EDUCATION

Carnegie Mellon University

Master of Science in Data Analytics for Science | QPA: 4.17/4.33

Pittsburgh, PA

May 2025

Vellore Institute of Technology

Bachelor of Technology in Computer Science and Engineering | CGPA: 8.88/10

Vellore, India

Jul 2024

SKILLS & CERTIFICATIONS

Machine Learning & AI: Scikit-learn, TensorFlow, Keras, XGBoost, Regression, Classification, Clustering, GANs, Synthetic Data, BERT, LLMs, RAG, ChromaDB, LLM Evaluation

Decision Science: Descriptive, Predictive & Prescriptive Analytics, Scenario Analysis, Monte Carlo Simulation, Optimization, Experimental Design, Hypothesis Testing, A/B Testing, Data Visualization

Data & Automation: Python, Pandas, SQL, PySpark, NumPy, GitHub Copilot Custom Agents

Product & Business: Problem Framing, Requirements Gathering, Stakeholder Collaboration, KPI & Metric Definition, Solution Design, Trade-Off Analysis, Data Storytelling, UAT, Project Ownership

Cloud & Platforms: Databricks, BigQuery, Oracle, GCP, Azure, Streamlit, Docker, CI/CD

Certification: AWS Certified Cloud Practitioner