

SHARANYA SHARMA she/her/hers

DATA ENGINEER

San Francisco/Santa Barbara, CA | (925) 699-2120 | sharanyasharma21@gmail.com
sharanyasharma.com | linkedin.com/in/sharanyasharma21 | github.com/sharanyasharma21

TECHNICAL SKILLS

Languages & Tools: Python, SQL, PySpark, Scala, JavaScript, Tableau

Data Engineering: ETL/ELT, dbt, Dagster, Airflow, Spark, Kafka, Hadoop, Apache Iceberg, DSPy

Warehouse & Databases: Databricks, Snowflake, Amazon Redshift, Microsoft Azure, BigQuery, OLAP/OLTP, Delta Lake
Cloud, DevOps, AI: AWS S3, GCP, Docker, Kubernetes, CI/CD, Git, GenAI/LLM, RAG, MCP

Machine Learning: Regression, Time Series Analysis, Regularization, Bayesian/Stochastic Modeling, XGBoost, LightGBM

EXPERIENCE

Scalable Commerce, Santa Barbara - CA

Data Engineer, Analytics

July 2024 - April 2026

- **ETL/ELT & Orchestration:** Built end-to-end data pipelines processing large-scale, multi-terabyte datasets using **PostgreSQL** and **Python** in **Redshift**, **S3**, **BigQuery**, and **Azure**. Modeled data transformations and implemented custom tests in dbt, while orchestrating production automation workflows with **Dagster**.
- Improved data pipeline SLAs by optimizing **dbt** incremental models, Redshift distribution/sort keys, partitioning, clustering, **Spark** tuning, and job parallelization, reducing warehouse costs, cutting daily pipeline runtime, and improving dashboard refresh capabilities.
- **Databricks Architecture:** Designed **Databricks lakehouse migration** POC to unify warehouse data fragmented across Redshift, BigQuery, Snowflake, and Azure.
- **AI & MCP:** Integrated a dbt **MCP server** and developed a custom **semantic layer** for business metrics across the data warehouse. Enabled non-technical teams to query trusted data and access standardized metrics more efficiently.
- **A/B Testing & Analytics:** Supported analytics for **250K+ monthly customers** and **\$50M+** in sales, covering product feature updates, conversion, acquisition, retention, ROI, cost-benefit analysis, hypothesis testing, and uplift measurement. Built custom dashboards using **Tableau**, Google Apps Script, and Google Analytics, with a strong focus on fraud and payment analytics tied to customer acquisition and retention.
- **Ad Hoc Analytics & Automation:** Managed high-priority ad hoc analytics requests across product, marketing, engineering, and stakeholder teams, translating complex business questions into solutions. Built custom internal systems and automation scripts for ETL pipeline alert monitoring, debugging, reporting, and workflow optimization. Independently supported product experiments, analytics initiatives, and operational reporting across 7 brands.

Stanford Linear Accelerator Laboratory (SLAC), Palo Alto - CA

Machine Learning Engineer

June 2024 - December 2024

- Developed deep learning models CNN, ResNet, and Vision Transformer to detect Bragg peaks in X-ray diffraction images, and determine protein crystal orientation through prediction of detector and beam tilts.
- Applied computational physics, Fourier optics, and diffraction analysis to reconstruct electron density maps for protein crystallography, for potential drug discovery research.

Appfolio, Santa Barbara - CA

Applied AI Engineer, Capstone at UCSB

Jan 2024 - May 2024

- Built an AI-powered data retrieval platform for property managers to query accounting documents, using LangChain, code interpreter workflows, few-shot prompting, and RAG with vector-database semantic search.
- Improved retrieval accuracy from 84% to 92%, developed a fuzzy-match evaluation framework to measure performance, and began exploring reinforcement learning to further optimize results.

TECHNICAL PROJECTS

Amazon - Product Intelligence on Databricks Lakehouse:

- Built a product intelligence lakehouse on Databricks using the Amazon Reviews 2023 dataset, ingesting large-scale product quality metadata into a **governed medallion (bronze/silver/gold) architecture**.
- Used **Lakeflow Spark** Declarative Pipelines, Structured **Streaming**, Auto Loader, **Delta Lake**, and **Unity Catalog** for pipeline processing, data governance, and duplication or outlier anomaly detection in data cleaning flow
- Implemented data quality expectations, simulated product metadata CDC, SLA/freshness monitoring, reconciliation checks, lineage-aware impact analysis, and **CI/CD** deployment with Databricks bundles and workflows.
- Added a **RAG**-based incident and product-intelligence layer using Databricks AI Search, reducing root-cause analysis time for pipeline failures and schema-drift issues by ~40% while surfacing downstream dashboard impact and recurring product-review themes for faster stakeholder resolution.

SnackShack - Feature Engineering Pipeline: Kubernetes-orchestrated data pipeline using Airflow and PySpark that ingests raw multi-source SnackShack orders data, normalizes it into model-training examples, writes Parquet datasets, and computes quality metrics.

EDUCATION & CERTIFICATIONS

University of California, Santa Barbara - Bachelor of Science, Statistics & Data Science + Technology Management
Databricks Associate Data Engineer (in progress)