# CHAITHANYA VEMURI

Junior Data Engineer | ML Engineer

Berlin, Germany | +49 15510944761 | chaithanyavemuri28@gmail.com

chaithanyavemuri.xyz | github.com/nikchey29 | linkedin.com/in/chaithanya-vemuri-141897264

Available from graduation (expected 01/2027) | Eligible on graduation for Germany's 18-month post-study residence permit (full-time work permitted)

## PROFILE

MSc Data Science student graduating in 01/2027. I build Python and SQL data pipelines and the machine learning services that run on them. My projects include a restartable pipeline over 411.7M e-commerce events, a recommender that beat its baseline by 46% on MRR@20, and a Kubernetes deployment on GCP with tested monitoring and rollback.

## TECHNICAL SKILLS

**Data engineering:** Python, SQL, DuckDB, PostgreSQL, Parquet, dbt, PySpark (Apache Spark), Redpanda (Kafka-compatible streaming).

**Machine learning:** PyTorch, scikit-learn, MLflow, time-series forecasting, recommender systems, offline evaluation.

**Cloud and DevOps:** GCP (GKE), Kubernetes, Docker, Terraform, Helm, Argo CD, GitHub Actions (CI/CD), Prometheus, Grafana.

**APIs and code quality:** FastAPI, pytest, Ruff, mypy, Git.

## PROJECTS

**REES46 V2 | Recommender System and Data Pipeline** github.com/nikchey29/rees46-v2

*Python | SQL | DuckDB | Parquet | FastAPI | PostgreSQL | MLflow | Docker*

- Processed **411.7M public e-commerce events** into Parquet with DuckDB, in bronze, silver and gold layers with validation and deduplication.
- Fixed a global aggregation that exhausted temporary storage after a ~30.6 GiB spill by redesigning it into monthly partitions and 64 session hash buckets, which made the pipeline restartable.
- Trained a ranking model on 1.50M interactions that **beat the popularity baseline by 46% on MRR@20** (0.02443 to 0.03556 on 5,000 held-out users), and served it through FastAPI with a popularity fallback.
- Set up GitHub Actions CI running pytest, Ruff and strict mypy, and containerized the service with Docker.

**OTTO | MSc Dissertation: Demand Forecasting** github.com/nikchey29/otto-demand-forecasting-dissertation

*Python | PyTorch | scikit-learn | Time Series | Chronological Validation*

- Streamed **216.7M OTTO click, cart and order events** in one pass into 672 hourly observations, and benchmarked eight 24-hour forecasting methods across three chronological folds and five neural seeds.
- Showed that a weekly seasonal baseline (**13.93% mean CV WAPE**) outperformed a PyTorch Transformer (17.67%). The selected model reached 9.80% cart WAPE and 11.34% order WAPE on a separate 96-hour holdout.

**FulfillAI | Cloud Deployment and Data Platform Lab** github.com/nikchey29/fulfillai

*GCP/GKE | Terraform | Kubernetes | Helm | GitHub Actions | Argo CD | dbt | PySpark*

- Provisioned GKE Autopilot, VPC and Artifact Registry with Terraform. Published images from GitHub Actions through Workload Identity Federation (no stored keys) and deployed pinned image digests with Helm and Argo CD.
- Verified API health and metrics in Prometheus and Grafana and tested Alertmanager-to-Slack routing. Diagnosed a deliberate ImagePullBackOff, rolled back, and verified Argo CD recovery.
- Built PostgreSQL and dbt pipelines for 50K synthetic orders and verified Redpanda and PySpark checkpoint recovery into PostgreSQL.

## EDUCATION

**Gisma University of Applied Sciences**, Berlin, Germany 08/2025 – 01/2027 (expected)

*MSc Data Science, AI & Digital Business*

Dissertation: Hourly cart and order forecasting from OTTO event logs.

**Sri Venkateswara College of Engineering**, Tirupati, India 10/2021 – 05/2025

*BTech Electronics & Communication Engineering* | Grade: 2.3 (German scale)

## LANGUAGES

**English:** professional working proficiency (IELTS Academic) **German:** A1, studying toward B1