Dante Secada-Oz

Email
contact@DanteOz.com

Website
DanteOz.com

RIT, Rochester Institute of Technology ’22 – B.S. Software Engineering

CORE SKILLS

Languages: Python, SQL, Java, C

Machine Learning

SKILLS: Natural Language Processing, Machine Learning, Time Series Forecasting, Information Retrieval, Model Inference, Model Evaluation

TOOLS: PyTorch, NumPy, JAX, HF Transformers, HF Accelerate, scikit-learn, XGBoost, LightGBM, ONNX, NVIDIA Triton

Data Engineering

SKILLS: Data Processing, Data Pipelining, Data Modeling (Dimensional/Relational), Data Quality, Data Analysis

TOOLS: Dagster, DBT, Pandas, Polars, DuckDB, Spark, Ray, Postgres, MySQL

Cloud Engineering

SKILLS: AWS, CI/CD, Infrastructure as Code, Containerization, Cloud Architecture

TOOLS: GitHub Actions, Docker, AWS CDK, Pulumi

SERVICES: S3, Kinesis, EMR, RDS, ECS, EC2, Lambda, CloudWatch, IoT, DynamoDB

EXPERIENCE

VizYourGov - Data Engineer (Platform)

2022 - Present

VizYourGov is a data-driven platform for visualizing the influence of money in U.S. politics.

- Led rewrite of data pipeline with 75K lines of Python, 200+ stored procedures, 250 tables/views, and 75 data sources.
- Refactored warehouse data model using dimensional data modeling.
- Developed LLM-based data cleaning tool, using retrieval augmented generation, few-shot prompting and chain-of-thought prompting.
- Improved web scraping performance by 50X using asyncio and distributed scraping pipeline.
- Built CI/CD and containerized deployment pipeline.
- Implemented structured logging, data catalog/lineage, data quality checks, and alerting.

Details →

Collibra

2020 - 2021

Two years of engineering co-op at the Data Intelligence Platform listed as the 7th most valuable data startup in the world.

Machine Learning Engineer (Business Process Automation)

2021

- Created and deployed a data pipeline to aggregate issues from engineering (Jira) and customer (Aha!) backlogs.
- Contextualized issues using customer and product metadata from various sources (Salesforce, Confluence, GitHub, etc).
- Redesigned ticket/feedback forms for ease of feature extraction.
- Created a classification model to automate issue allocation and prioritization.

Details →

Machine Learning Engineer (Knowledge Graph)

2020

- Conducted tabular representation-learning research for entity deduplication, with all data remaining on edge for client privacy and security.
- Conducted data mining research on whether corporate investment in data leads to significant ROI (for Collibra-partner UC San Diego BlockLAB).
- Implemented ETL pipeline, transforming unstructured corpora of academic and business journals into a knowledge graph, using active learning, clustering, and topic modeling.

Details →

PROJECTS

FastSearch

End-to-end semantic search engine indexing ~300-hour machine learning video lectures for **fast.ai** course.

- Fine-tuned bi- and cross-encoder models using cross-architecture knowledge distillation.
- Optimized models for low-latency retrieval and ranking.
- Collected, filtered and labeled dataset of ~1,000 **fast.ai** questions and ~27,000 lecture segments.
- Built data pipeline to scrape and transcribe new video lectures and incrementally update ANN search index.
- Instrumented user query and result feedback logging for model retraining.

- Deployed using CI/CD pipeline and IAC best practices.
- Built MLOps pipeline to redeploy from model registry and backfill ANN index.

FastSearch website → Write-up → GitHub code →

Lockheed Martin - IoT Pipeline and Anomaly Detection (RIT Senior Project)

Real time dashboard to monitor health and performance of factory machinery at the world's largest military/aerospace company.

- Built streaming data pipeline for factory machine telemetry from IoT sensors.
- Trained time series based anomaly detection model to predict machine failures.
- Deploy pipeline and dashboard to AWS: pipeline (Kinesis, SQS, Greengrass, MTConnect), hosting (Amplify), server-less backend (Lambda, API Gateway), NoSQL databases (DynamoDB, Neptune) and IAC (AWS CDK).

Details →

JetBrains Research – Model deployment (Software Engineering Research)

JetBrains IntelliJ IDE plugin which lints Java code base for code snippets to refactor.

- Developed random forest and XGBoost models to lint Java method for refactoring.
- Deployed TensorFlow, scikit-learn and XGBoost in Java based JetBrains IntelliJ IDE plugin.
- Validated data collection and model validation scheme using permutation feature importance and adversarial validation.

Details →