JONATHAN ESPERANZA

jonathanesperanza67@gmail.com | linkedin/jonathanesperanza | github/jon-esperanza

EXPERIENCE

Credit Karma Charlotte, NC

Scala, Python, Tensorflow, Apache Spark, Apache Beam, Airflow, Kubernetes, Terraform, Google Cloud Platform

Software Engineer III, Offline Recommendations Platform *September 2024 - Present*

- Proactively engaged with data scientists to understand their pain points and challenges in the model lifecycle, leading to a collaborative relationship, enhanced tooling, and better architectural decisions for production data pipelines and inference.
- Introduced customizable post-processing for model predictions to enable data scientists to modify behavior without engineering involvement, streamlining business initiatives and accelerating model agility.
- Successfully modernized our online ML architecture using NVIDIA Triton Inference Server resulting in reduced latency, lower CPU utilization, and large cost savings.

Software Engineer II, Offline Recommendations Platform *February 2023 - September 2024*

- Responsible for driving 1.5B personalized notifications per month, attributing to 24% of sitewide traffic through machine learning (ML) pipelines.
- Owned and deployed several ML and data features through cross-functional efforts with marketing, product, and data science teams resulting in notable lifts to engagement and revenue.
- Achieved a remarkable 70% cost reduction on Cloud Dataflow, saving $1.8 million annually, by optimizing performance for large, inefficient data joins in our batch inference ML pipelines.
- Designed, built, and delivered a pre-production environment, significantly reducing outages associated with major feature releases by 95% and unlocking performance benchmarks and tuning.
- Drove efforts for platform re-architecture initiative to adopt Airflow, Terraform, and inference graphs at scale leading to enhanced infrastructure stability, flexible ML feature scalability, and enabling self-serve platform.

Software Engineer I, Recommendations: Prediction Services *August 2022 - February 2023*

- Contributed to scaling, maintaining, and operating critical model serving and feature store services to serve recommendations to 150M+ members in real time.
- Assisted in the effort to migrate our Tensorflow model scoring service from TF 1.x to TF 2.x, ensuring a smooth transition and improved latency/performance.
- Collaborated on a proof-of-concept redesign of a critical pipeline in our recommendation request flow logging to reduce storage costs and accelerate data delivery to analytics.

EDUCATION

University of North Carolina in Charlotte Charlotte, NC

B.S Computer Science, Software Engineering *August 2022*

SKILLS

Programming Languages	Java, Scala, Python, SQL
Cloud	Google Cloud Platform (Cloud Storage, BigQuery, BigTable, Cloud Dataflow, Cloud Functions, Pub/Sub); Experience with AWS and Microsoft Azure
Technologies and Frameworks	Apache Beam, Tensorflow, Apache Spark, Kubernetes, Docker, Terraform, Airflow
Development Tools	Git, CircleCI, GitLab, GitHub