Min Gu (Min) Jo

510-365-4988 | mgj9993@gmail.com | http://mingujo.github.io | linkedin.com/in/mingujo/

SKILLS

Languages: Python, Go, Scala, SQL, Java
Frameworks & Libraries: Apache Spark, Hadoop, Hive, Kafka, Airflow, Kubernetes, Docker, AWS, Hashicorp (Terraform, Vault), Snowflake, TensorFlow, Redis
Technical Concentrations: Data Engineering • ML Platform • Distributed data processing • Streaming • Data Warehousing • Data Modeling • Distributed System • Performance Tuning • ML • NLP

EDUCATION

University of California, Berkeley Class of 2016
B.A. in Computer Science, Statistics, and Economics

WORK EXPERIENCE

Software Engineer | Opendoor, San Francisco CA Mar 2018 - Present

- Developed in-house distributed data processing and scheduling system (Spark on Kubernetes) which serves 100+ engineers and processes ~1TB of real estate data on daily basis
- Led and executed Spark infrastructure migration from in-house Kubernetes to managed service, Amazon EMR
  - Reduced fixed OPEX cost on data engineering by 45%
  - Enhanced reliability of Spark batch processing jobs by 68%
  - "*Using Amazon EMR to build a Spark ecosystem at Opendoor*" Presentation given at AWS re:Invent 2019 youtube.com/watch?v=PkoYqp_6WTc
- Built in-house observer system to regularly validate quality of ingested data and alert for SLA violation
- Improved usability of Opendoor data lake (S3) by installing Hive Metastore

Software Engineer | Leadgenius, Berkeley CA Jan 2017 - Feb 2018

- Built an automatic sales outreach email reply labeling ML application to serve 50+ customers
  - Classified email replies in 7 different labels by applying NLP (LSTM) algorithm using Tensorflow
  - Alerted customers of positive replies in real-time by deploying a trained model on a server
    ⇨ Overall Accuracy: 92.8% (Accuracy on "positive" reply: 89.2%)
    ⇨ Increased returning visitor rate of the company's outreach product by 72%
- Developed an ETL processing pipeline to store 25+ mil U.S. based company and 30+ mil professional data from a variety of sources
- Indexed and stored transformed data by designing data deduplication algorithm
- Implemented distributed search for database of 25+ mil company data using ElasticSearch

Research Assistant | Berkeley Institute of Data Science, Berkeley CA Jan 2016 – Dec 2016

- Implemented data ingestion pipeline of 5000+ movie review data into S3 via web scraping framework
- Trained binary sentimental classification model to label user reviews using Bag of Words model

PROJECT EXPERIENCE

UC Berkeley Family Housing - Open House Scheduling Calendar Fall 2016

- Developed a calendar web app using Rails to serve 15+ UCB housing staffs for coordinating open house schedules with 30+ resident assistants. Automated email notification for any schedule changes.

Kaggle Challenge: Rossmann Drugstore Store Sales Prediction Spring 2016

- Used and compared 3 different machine learning algorithms to forecast drugstore daily sales: multivariate linear regression, random forest regression, and gradient boosting with regression trees