Min Gu (Min) Jo

510-365-4988 | mgj9993@gmail.com | http://mingujo.github.io | 2225 Channing way, Berkeley CA 94704

SKILLS

Languages: Python, Go, Scala, SQL, Java
Frameworks & Libraries: Apache Spark, Hive, Airflow, HDFS, Kubernetes, AWS(EMR, DynamoDB, Terraform, RDS), Snowflake, Bigquery, Concourse, TensorFlow, Redis, Celery, ElasticSearch
Technical Concentrations: ETL • Distributed data processing • Streaming • Data Warehousing • SQL • NoSQL SOA • Hadoop Ecosystem • Analytics • 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 batch data processing and scheduling system (Airflow + Spark + Kubernetes) which serves 100+ Opendoor engineers and processes ~1000TB of real estate data daily
- Led and executed Spark cluster migration from in-house Kubernetes to AWS EMR
  - Reduced fixed platform cost on data engineering by 38%
  - Enhanced reliability of Spark streaming and batch jobs by 20%
  - Expedited batch processing jobs by 40%
- Built in-house observer system to regularly validate quality of ingested data and alert for freshness SLA violation
- Improved usability of Opendoor data lake (S3) by installing external Hive Metastore

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

- Built an automatic sales outreach email reply labeling application to serve 50+ customers from scratch
  - Classified email replies in 7 different labels by applying NLP algorithms
  - 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 merged 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