Yu-Wen (Sonic) Lai

yuwen.lai@nyu.edu | (929) 401-8800 | Seattle, WA | linkedin.com/in/sonicywlai | github.com/hsnusonic

Profile

Innovative software engineer with extensive exposure to distributed systems and back-end development. Passionate about data engineering and large-scale systems. Proficient in problem solving and collaborating with cross-disciplinary people.

Work Experience

Cloudera Palo Alto, CA

Software Engineer Intern, Hive *Jun 2020 - Aug 2020*

- Revised a config for skipping trash folder and contributed to the open source community
- Diagnosed the performance and memory footprint of HyperLogLog algorithm and researched alternative ways to improve
- Discovered a performance bottleneck by using CPU Flame Graph, reaching 15x improvement in elapsed time for query optimization
- Enhanced Hive's query federation for MSSQL by testing and resolving a sorting pushdown issue

Acer New Taipei, TW

Project Engineer, Value Lab *Sep 2016 - Jul 2019*

- **Cancer Immunotherapy**: **Distributed Computing, Python, Bash**
  - Planned the computing infrastructure by Ansible and Sun Grid Engine queuing system
  - Implemented and scaled DNA data analysis pipeline by Workflow Description Language, enhancing efficiency up to 10x by taking advantage of parallel processing and better memory allocation
  - Standardized package management on servers by Conda, decreasing colleagues' 85% time for building a testing environment
- **Big Data Platform Building for FarEasTone Telecommunications**: **Spark, Kafka, HBase**
  - Collaborated with IBM to design real-time analysis of location-based user data, gaining $645K contract for Acer
  - Achieved processing 1.5M records/min by PySpark program which extracts data from Kafka and loads data into HBase
- **Smart Taxi Operating Platform for Taiwan Taxi**: **Spark, Amazon Web Services, Hadoop**
  - Established RESTful data collection APIs using Flask and Nginx for the demand prediction APP used by thousands of taxi drivers
  - Developed web crawlers for collecting information about concerts and exhibitions, improving 5% accuracy of the prediction model
  - Analyzed taxi cabs in real-time by Spark Streaming, providing up-to-date feedback on the prediction
  - Built the data pipeline from public cloud to on-prem cluster by Fluentd, Sqoop, Hive, and Impala, importing 35 GB data daily

Projects

- **A miniature relational database,** New York University: **Python, Unit Testing** *Dec 2019*
  - Implemented indexing by in memory B-trees and hash table, which supports operations of selection and join.
  - Constructed test cases to cover edge cases by Python's unit testing framework - unittest
- **Chinese Characters Recognition,** National Taiwan University: **Python, Machine Learning** *Jan 2015*
  - Attained 70% precision for 32 possible labels of handwritten Chinese characters by KNN, random forest, and SVM
  - Programmed data pre-processing, training, and validation by Numpy, scikit-learn, and LIBSVM

Education

New York University New York, NY

Master of Science in Computer Science, GPA: 3.9/4.0 *Expected Graduation: Dec 2020*

Coursework: Programming Languages, Fundamental Algorithms, OS, Database Systems, NLP, Computer Security

National Taiwan University Taipei, TW

Master of Science in Physics *Jun 2016*

Bachelor of Science in Physics *Jun 2012*

Coursework: OOP, GPU Programming, Computer Architecture, Machine Learning, Information Retrieval

Skills

- **Languages**: Python, Java, C/C++, Scala, SQL, Bash
- **Libraries**: PySpark, Scrapy, Flask, Numpy, Pandas, scikit-learn, TensorFlow, CUDA
- **Tools**: Git, MySQL, Impala, Hive, Sqoop, Kafka, Docker, Ansible, Hadoop, HBase, Google Cloud Platform, Amazon Web Service, Azure