Tianwen Chu

Rockville, MD - Email me on Indeed: indeed.com/r/Tianwen-Chu/9dc3e79f2f4457b2

I am seeking a data scientist/software engineer full time role to embrace more challenge than from current company. I have a solid fundamental of machine learning, algorithm and coding skills. I have used spark and scala for more than two years which out stands myself in the market.
Willing to relocate: Anywhere
Sponsorship required to work in the US

WORK EXPERIENCE

Lead Data Scientist

Fed Centric Technologies - College Park, MD - September 2015 to Present

Worked with two Phds on a complex proof of concept in collaboration with the National Cancer Institute/ Frederick National Laboratory for Cancer Research/ Leidos Biomedical Research. The POC is about applying spectral clustering on genomic variations among individuals from 1000 Genome Project Data. Wrote cpp program to do query from a graph database (sparksee) with multi-threading and used python to implement the clustering algorithm with numpy. This work won 3rd place out of 42 at the 4th Annual BioMedical Informatics Symposium at Georgetown University on October 16, 2015 and 1st place out of 78 at the Bio-IT World Conference Expo in Boston, MA on April 7, 2016.

Wrote a tax fraud checking application with full stack development. Used sparksee as the database storing reference W2s and 1099s, then matched incoming persons' 1040 forms by operating graph queries. Wrote php and python program as backend to process xml input data, javascripts as front end web UI for functionality and visualization of process and check results.

Led team of 3 work on a bio-science paper reproduce. Wrote a cpp program to parse virus RNA sequence into k-mers, mapping k-mers to consecutive amino-acids and finally convert original string sequences into mathematic vectors in multi-threading. Applying supervised SVM in matlab to classify multiple virus types and cross-validation to measure test accuracy.

On a 192 core scale-up sgi server machine, wrote

scala program to ingest genomic VCF file from 1000 Genome Project to backend with spark. Converting large VCF file into Dataset data structure in spark after a series of transformation. Used MemSQL as database for spark to do ingestion and query. Compared performances of ingestion between from spark to memsql, spark to HDFS parquet and spark to ignite. Wrote benchmark scripts to measure the performance of ingestion by varing different spark configuration parameters. Also compared the ingestion and query performance with RDF graph from Cray. Worked in team of 4 on a benchmark project of running whole genomic pipeline on SGI machine.

Wrote Q script in scala to run step from aligning small sequences to reference genome until variant calling and calibration in map-reduce way.

EDUCATION

Master of Science in Computer Science, Robotics, AI

Carnegie Mellon University(GPA 3.7) - Pittsburgh, PA
August 2013 to May 2015

Bachelor of Mechanical and Automation Engi in Mechanical Engineering

Southeast University - Nanjing, CN
September 2009 to June 2013

SKILLS

MATLAB (3 years), PHP (1 year), MEMSQL (1 year), CSS (Less than 1 year), Tensor Flow (1 year), Scala (2 years), C++ (3 years), Java (2 years), apache spark (2 years), Python (3 years), machine learning (2 years), algorithm (2 years), Computer Vision (2 years), Linux (3 years), SQL (2 years)

LINKS

http://linkedin.com/in/tianwen-chu-72673180