DB TSAI

E-mail: dbtsai@dbtsai.com *Web: www.dbtsai.com* *LinkedIn: www.linkedin.com/in/dbtsai* *GitHub: www.github.com/dbtsai*

Summary

- I specialize in big data machine learning with strong background in theoretical statistics and mathematics.
- I've implemented various distributed machine learning algorithms using Hadoop and Spark for large-scale data processing, and contributed back to open source communities.
- I've been actively involved with the open source Apache Spark developement as a contributor.

Specialties

- Distributed Machine Learning and Data Mining.
- Apache Hadoop and Spark stack.
- Computer languages such as Scala, Java, Python, C, and C++.
- Mathematical scripting languages (Matlab and R).
- Parallel Computing and Big Data Processing using MapReduce and MPI.

Experience

- **Apache Spark** — A fast and general engine for large-scale data processing
  Committer *May 2015 to current*
  - My contributions, https://github.com/apache/spark/commits/master?author=dbtsai
  - Implmented new features such as L-BFGS, and Logistic Regression, etc.
  - Built test infrastructure in MLlib for comparing floating-point numbers using implicit conversion.
  - Improved MLlib performance such that K-mean runs 3x faster, and StandardScaler runs 10x faster.
  - Fixed various bugs and wrote documentation.
- **Netflix, Los Gatos, CA** — A Leading Provider of Internet Streaming Media Available Worldwide
  Senior Research Engineer *April. 2015 to current*
  - Spark Machine Learning Training Pipeline.
- **SF Machine Learning Meetup, CA** — People with Shared Interests of Machine Learning and Big Data
  Co-Organizer *Jun. 2013 to current*
  - http://www.meetup.com/sfmachinelearning/
  - Had more than 2700 machine learning enthusiasts in the community.
  - Hold the meetup monthly, and invited famous speakers in industry and academic to give talks.
- **Alpine Data Labs, San Francisco, CA** — The Leader in Data Science for Big Data
  Machine Learning Lead *Aug. 2014 to April 2015*
  Machine Learning Engineer *Apr. 2013 to Aug. 2014*
  - Developed scalable Multinomial Logistic Regression and Linear Regression with elastic-net regularization which linearly combines the L1 and L2 penalties in Apache Spark. Implemented OWLQN for L1/L2 regularized optimization.
  - Developed scalable algorithms such as Decision Tree, Variable Selection based on Information Gain, exact one-pass Linear Regression with L2 penalty, and PCA in Hadoop MapReduce.
  - Migrated build infrastructure from ANT to SBT for better third party library dependency management using the Maven central repository, better intergation with Jenkins for continuous integration, better developement/debuging experience for developers, and easier release build.
- **KeeKa, StartX 2012 Summer, Stanford, CA** — A Social Network Connecting People through Fashion
  Co-founder and CTO *Jan. 2012 to Mar. 2013*
  - Planned the strategies and invented a disruptive product.
  - Designed the architecture of the website, including deployment, front-end, and back-end systems.
  - Coordinated the designer, front-end team, and back-end team and performed the code review to ensure reliability, effectiveness, progress, and productivity.

Recent Talks and Book (My slides can be found on: http://www.slideshare.net/dbtsai/)

- **Lambda Architecture with Apache Spark**, Galvanize, San Francisco, CA
  Next.ML Conference *Jan. 17, 2015*
- **Large-Scale Machine Learning with Apache Spark**, Moscone Center, San Francisco, CA
  Internet of Things Conference *Oct. 20, 2014*
- **Alpine Invovation to Spark**, Cloudera, Palo Alto, CA
  Cloudera & Alpine Data Labs tech talks *Aug. 14, 2014*
- **Multinomial Logistic Regression with Apache Spark**, Hacker Dojo, Mountain View, CA
  Silicon Valley Machine Learning Meetup *June 20, 2014*
- **Multinomial Logistic Regression with Apache Spark**, Alpine Data Labs, San Francisco, CA
  SF Machine Learning Meetup *May 1, 2014*
- **A Guide to Having Fun with the Next Generation Linux, Ubuntu by S.-W. Lee and D.-B. Tsai**, Taipei
  ISBN: 9867199979, GrandTech Press. Among the top 5 best-selling computer-science books from Nov. 2006 to Jan. 2007 in the Chinese book market in Taiwan. *Sept. 14, 2006*

Education

- **Stanford University**, California, U.S.A.
  ABD in Applied Physics Ph.D. program *Sept. 2010 to June 2012*
  M.S. in Electrical Engineering *Sept. 2010 to June 2012*
- **National Taiwan University**, Taipei, Taiwan
  M.S. in Physics *Sept. 2006 to July 2008*
- **National Cheng Kung University**, Tainan, Taiwan
  B.S. in Physics *Sept. 2002 to June 2006*