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 have implemented various distributed machine learning algorithms using Hadoop and Spark for large-scale data processing, and contributed back to open source communities.
- I have been actively involved with the open source Apache Spark developement as a committer.

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
  - Implemented new features such as L-BFGS, and Multinomial / Binomial Logistic Regression, etc.
  - Conducted code review for other contributors, and guided them until the code is merged.
  - Fixed various bugs, wrote documentation and performed performance optimization.
- **Netflix, Los Gatos, CA** — A Leading Provider of Internet Streaming Media Available Worldwide
  Senior Research Engineer *April. 2015 to current*
  - Worked on personalized recommendation algorithms and machine learning infrastructure.
  - Architected and implemented Distributed Time Travel Machine for Feature Generation using Apache Spark, which enables our researchers to quickly try ideas for new features on historical data such that running offline experiments and transitioning to online A/B tests is seamless. This framework reduces the time to bring an offline experiments to online A/B tests from months to weeks, and significantly removes the offline/online discrepancy because of sharing the feature generation logics between offline/online. U.S. Patent filed February 2016. Patent Pending.
  - Implemented categorical feature learner in Netflix's in-house GBDT (Gradient Boosting Decision Tree) implementation as part of the global algorithm effort to incorporate the country and language categorical signals.
  - Implemented Weighted Logistic Regression in open source Apache Spark ML which is used in Netflix's personalized page algorithms for constructing the rows in the homepage.
  - Worked closely with Apache Spark community to merge our changes, and implemented new features for our needs.
- **SF Machine Learning Meetup, CA** — People with Shared Interests of Machine Learning and Big Data
  Co-Organizer *Jun. 2013 to July. 2015*
  - 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
  Lead Machine Learning Engineer *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.

Publications (https://www.dbtsai.com/publications/)

- **MLlib: Machine Learning in Apache Spark**,
  Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, **DB Tsai**, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar
  Journal of Machine Learning Research, Vol 17 (34) pp. 1-7 *April 2016*
- **Distributed Time Travel for Feature Generation**,
  Hossein Taghavi, Prasanna Padmanabhan, **DB Tsai**, Faisal Zakaria Siddiqi, and Justin Basilico
  US Patent Pending *Feb. 2016*
- **Quantum Zeno and anti-Zeno effect of nanomechanical resonator measured by a point contact**,
  Po-Wen Chen, **DB Tsai**, and Philip Bennett
  Physical Review B 81, 115307 *March 2010*
- **Optimal control of the silicon-based donor-electron-spin quantum computing**,
  DB Tsai, Po-Wen Chen, and Hsi-Sheng Goan
  Physical Review A 79, 060306(R) *June 2009*
- **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*

Talks (https://www.dbtsai.com/talks/)

- **Distributed Time Travel for Feature Generation**, Yelp, San Francisco, CA
  SF Big Analytics *Mar. 24, 2016*
- **Distributed Time Travel for Feature Generation**, Hilton Midtown, New York, NY
  Spark Summit *Feb. 17, 2016*
- **Large-Scale Elastic-Net Regularized Generalized Linear Models**, The Hilton Union Square, San Francisco, CA
  Spark Summit *June 15, 2015*
- **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*

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*