## Contact

www.linkedin.com/in/dbtsai (LinkedIn)
www.dbtsai.com (Personal)
github.com/dbtsai (Other)

## Top Skills

Machine Learning

Hadoop

Linux

## Languages

English (Full Professional)

Chinese (Native or Bilingual)

## Publications

A Guide to Having Fun with the Next Generation Linux, Ubuntu (in Chinese)

MLlib: Machine Learning in Apache Spark

A quantum effect in the classical limit: nonequilibrium tunneling in the Duffing oscillator

Optimal control of the silicon-based donor-electron-spin quantum computing

Quantum Zeno and anti-Zeno effect of nanomechanical resonator measured by a point contact

## Patents

Distributed Time Travel for Feature Generation

# DB Tsai

Senior Engineering Manager at Databricks
Cupertino, California, United States

## Summary

As an engineering leader at the Apple Data Platform, I possess the capability to drive projects that are undefined and ambiguous from their initial stages to successful production. With a solid technical background as a member of the Project Management Committee (PMC) and a committer for various Apache projects like Spark, Yuiknorn, and SystemDS, I have the expertise to attract and recruit top talents globally while fostering a strong engineering culture within my teams.

I have achieved notable accomplishments by building and leading multiple teams at Apple, including those dedicated to Spark, Flink, and Data Security. The collective efforts of these teams were recognized and honored with the prestigious ACM SIGMOD Systems award in both 2023 and 2022, marking two consecutive years of achievement. Many projects developed within my teams have also attained industry-standard status, enabling wider accessibility and utilization among diverse user communities.

Through my leadership and technical acumen, I have consistently driven innovation, established high-performing teams, and contributed to the advancement of the Apple Data Platform, resulting in accolades and the widespread adoption of our projects throughout Apple and the industry.

Talks: https://www.dbtsai.com/talks
Publications: https://www.dbtsai.com/publications

## Experience

### Databricks

Senior Engineering Manager
November 2024 - Present (3 months)
Mountain View, California, United States

## The Apache Software Foundation

9 years 8 months

### Apache YuniKorn PMC member / Committer

March 2019 - Present (5 years 11 months)

### Apache Spark PMC Member / Committer

June 2015 - Present (9 years 8 months)

– My contributions, https://github.com/apache/spark/commits/master?author=dbtsai
– Designed and implemented L-BFGS, Multinomial / Binomial Logistic Regression in Spark, and several other features.

## Apple

6 years 7 months

### Head of Core Data Platform

October 2023 - October 2024 (1 year 1 month)
Cupertino, California, United States

I lead three teams at Apple: Spark, Flink, and Data Security, growing them from 2 to 25 members. It's an honor to lead these award-winning teams, recognized with the ACM SIGMOD Award in 2022 and 2023, demonstrating our impactful work and leadership in the big data industry.

1. Our Spark team at Apple is exceptionally strong, with several Spark PMC members and committers. Many members co-received the 2022 ACM SIGMOD System Award for their contributions to Apache Spark. We leverage our open-source influence to align developments with Apple's needs. The team has open-sourced Apache DataFusion Comet, a Spark native accelerator that accelerates data insights, improving business outcomes and significantly reducing compute expenses, saving millions each month for Apple.

2. The Flink team at Apple includes several Flink PMC members and committers, with two co-receiving the 2023 ACM SIGMOD System Award for their contributions to Apache Flink. They developed a new Java-based Flink operator tailored to Apple's unique use cases and open-sourced it under Apache Flink. Recently, Google retired their Flink K8s operator and now recommends ours. It's incredible to see our project's adoption beyond Apple.

3. At Apple, safeguarding customer data is paramount. Our security team developed Parquet Modular Encryption and Apache Iceberg Table Encryption, providing scalable columnar data encryption compatible with query engines

like Spark, Trino, Flink, and Hive. This ensures the protection of vast amounts of sensitive data, spanning multiple petabytes. We've open-sourced this technology, and its widespread adoption as an industry standard has been particularly gratifying, enabling broader accessibility and utilization.

### Senior Engineering Manager

October 2019 - September 2023 (4 years)

Cupertino, California, United States

### Staff Software Engineer, Data Platform

April 2018 - September 2019 (1 year 6 months)

Cupertino, California, United States

Responsible for creating Apple Data Platform strategies, innovating Spark functionalities for Apple's needs, and significantly contributing to Apache Spark's development. As an Apache Spark PMC member, advocated for Apple's requirements in the broader community.

## Netflix

### Machine Learning Researcher / Technical Lead

April 2015 - March 2018 (3 years)

Los Gatos, CA

– Lead and architect the personalized recommendation pipelines and machine learning infrastructure using Apache Spark.
– Architect and implement Distributed Time Travel Machine for Feature Generation using Apache Spark, which enables our researchers to quickly try ideas with 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.
– Implement 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 for global launch.
– Implement 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.
– Work closely with Apache Spark community to merge our changes, and implement new features for our needs.

## Alpine Data

Senior Machine Learning Engineer
April 2013 - April 2015 (2 years 1 month)
San Francisco, CA

– 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 integration with Jenkins for continuous integration, better developement/ debuging experience for developers, and easier release build.

DuJour
Co-founder
January 2012 - March 2013 (1 year 3 months)
Palo Alto, CA

## Education

Stanford University
Doctor of Philosophy (Ph.D.) Program, Applied Physics · (2010 - 2012)

Stanford University
Master's degree, Electrical Engineering · (2010 - 2012)

National Taiwan University
Master's degree, Physics · (2006 - 2008)

National Cheng Kung University
Bachelor's degree, Physics · (2002 - 2006)