CHENGHAO LYU

chenghao@cs.umass.edu — https://chenghao.pages.dev — LinkedIn

EDUCATION

University of Massachusetts Amherst, Amherst, MA 09/2018 — 06/2024
PhD in Computer Science (Advisors: Yanlei Diao and Prashant Shenoy)
MS in Computer Science (GPA: 3.94)
Fudan University, Shanghai, China 09/2011 — 06/2018
MS in Computer Science (Advisor: X. Sean Wang)
BS in Electronic Engineering (Rank: 10/100)

RESEARCH INTERESTS

My research lies in the intersection of big data analytics systems, machine learning, and multi-objective optimization, with a focus on designing optimizers to auto-configure parameters in large-scale systems to achieve improved performance and cost reduction. I apply machine learning to construct performance models within a multi-channel input framework. These optimizers have demonstrated their effectiveness in Apache Spark, particularly in Spark SQL, and Alibaba MaxCompute.

PROFESSIONAL EXPERIENCE

CEDAR Group at Ecole Polytechnique, Paris, France 10/2021 — 02/2024
Scientific Collaborator (Mentor: Yanlei Diao)

- Work for the European Research Council (ERC) project at CEDAR, a joint team between Inria Saclay and LIX.
- Design the auto-tuning strategy for Spark parameters both at compile time and runtime to reduce latency and cost.
- Recompile Spark to enable the injection of customized rules for automatic configuration of SQL parameters at runtime.
- Collect traces and build performance models for queries and stages via machine learning.
- Improved latency and cost have been achieved by applying our optimizer to TPCH and TPCDS.

Alibaba DAMO Academy, Hangzhou, China 02/2020 — 09/2021
Research Intern (Mentors: Kai Zeng & Yaliang Li)

- Worked with the scheduling team at AliCloud to solve a resource optimization (RO) problem in Alibaba MaxCompute.
- Presented a MaxCompute-based big data system that supports multi-objective RO towards a stage, the scheduling unit that involves instances running in parallel, via fine-grained instance-level modeling and optimization.
- Derived a latency-aware placement plan to reduce the stage latency and instance-specific resource assignment plans to further reduce latency and cost in a hierarchical multi-objective optimization framework, with optimality proofs.
- Developed an ETL pipeline that pulls data from five sources to learn fine-grained instance-level models.
- Reduced 36-37% latency and 37-75% cost over three production workloads compared to Alibaba's default scheduler.
- One paper was accepted in VLDB'22.

DREAMLab at University of Massachusetts Amherst, Amherst, MA, USA 01/2019 — 08/2021
Research Assistant (Advisors: Yanlei Diao & Prashant Shenoy)

- Worked on a Unified Data Analytics Optimizer (UDAO) that auto-configures Spark parameters to meet multiple task objectives (latency, throughput, cost, etc.) with performance modeling and multi-objective optimization (MOO).
- Set up TPCxBB benchmark in a self-maintained Spark cluster and collected 20K traces of task-configuration pairs.
- Disentangled the task embedding in a low-dimension space by feeding the observed traces as the input to an autoencoder with a customized triplet loss. Build performance models from the task embedding and the configuration.
- Designed a multi-objective gradient-based solver that helps UDAO achieve a 2-50x speedup over existing MOO methods with good coverage of the Pareto frontier. Compared to SOTAs, UDAO's recommended configurations yield a 26-49% reduction of the TPCxBB benchmark running time while adapting to different user preferences on multiple objectives.
- Led the code release.
- Two papers were accepted in VLDB'19 and ICDE'21.

SELECTED PUBLICATIONS

- **Fine-Grained Modeling and Optimization for Intelligent Resource Management in Big Data Processing**
  Chenghao Lyu, Qi Fan, Fei Song, Arnab Sinha, Yanlei Diao, Wei Chen, Li Ma, Yihui Feng, Yaliang Li, Kai Zeng, Jingren Zhou
  In Proceedings of the VLDB Endowment, Volume 15, Issue 11, pp 3098–3111.
- **Spark-based Cloud Data Analytics using Multi-Objective Optimization**
  Fei Song*, Khaled Zaouk, **Chenghao Lyu***, Arnab Sinha, Qi Fan, Yanlei Diao, Prashant Shenoy (*equal contribution)
  In 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 2021 pp. 396-407
- **UDAO: A Next-Generation Unified Data Analytics Optimizer**
  Khaled Zaouk, Fei Song, **Chenghao Lyu**, Yanlei Diao
  In Proceedings of the VLDB Endowment, Volume 12, Issue 12, pp 1934–1937.

TEACHING EXPERIENCE

Teaching Assistant at University of Massachusetts Amherst

- COMPSCI 514 Algorithm for Data Science (2023 Spring)
- COMPSCI 645 Databases (2020 Spring)
- COMPSCI 105 Computer Literacy (2018 Fall)

Teaching Assistant at Fudan University

- SOFT130039.01 Discrete Mathematics (2016 Fall)

SKILLS

- **Relevant Coursework:** Neural Network (A), Reinforcement Learning (A), Machine Learning (A-), Optimization (A), Advanced Algorithms (A), Databases Design (A), Distributed System (A).
- **Programming:** Experienced in Python (Pytorch); familiar with Scala, Java, C++., etc.
- **Software:** Proficient in Apache Spark; familiar with Apache Hadoop, Flink, Kafka, etc.

SELECTED HONORS AND AWARDS

Academic

- Tung OOCL Scholarship (First Prize Scholarship) at Fudan University
- Scholarship of Academic Excellence at Fudan University
- 3× Outstanding student at Fudan University
- Honorable Mention in 2014 Mathematical Contest in Modeling
- Entrance examination waived for graduate education

Sports

- 1× Championship in Fudan Basketball Cup, 3× Championship in Fudan Basketball League, 1× first scoring player in sub-campus Basketball Cup
- The former member of the Fudan Dragon Boat Team
- 1^{st} Prize in the 2015 Dragon Boat Competition compared with C9 Colleges in China
- Advanced Individual in Sports at Fudan University