Hayes Zhao

hayes.zhao09@gmail.com • 703-915-8580 • www.linkedin.com/in/hayes-zhao • https://hzsep.github.io • Sunnyvale, CA & Herndon,VA

Data Engineer with a strong foundation in AI infrastructure, distributed computing, and system optimization. Proven ability to build distributed pipelines and optimization models to improve infrastructure utilization, reduce operational overhead, and scale AI workloads, aligning with requirements for AI/ML and infrastructure roles.

Experience

Data Engineer **| Bytedance** Ashburn, VA | Sep. 2025 – Present

Global ETL Pipeline (Business and Finance Data base)

- Built a production ETL pipeline integrating business and finance databases for global Data Center operations, improving data consistency from 98.0% to 99.9%, ensuring reliability through robust data processing and validation.
- Orchestrated daily ETL jobs with Apache Airflow using Python scripts to process 1M assets and 10K incremental updates, implementing retry logic and SLA monitoring to maintain near-real-time data availability.
- Scaled transformations and reconciliation logic with Spark, reducing manual triage by 90% through automated validation checks and improving accuracy in data center operations.
- Partnered with finance and operations teams to prioritize data requirements and streamline runbooks, improving incident response and operational transparency.

Capacity & Compute Planning

- Developed a Python-based planning system combining a MILP solver with reinforcement learning policies to optimize procurement, placement, and migration under power and lead-time constraints, accelerating plan generation and reducing procurement cost.
- Modeled downtime and migration costs for an internal fleet, selecting plans that maximize ROI while meeting availability requirements for large-scale ML workloads.
- Delivered +6% average compute uplift at fixed power via optimized allocation and migration schedules; built offline evaluation and guardrails to validate plans and ensure reliability.

Research Assistant **| George Washington University** Washington, DC | Mar. 2024 – Sep. 2024

- Built a topology-aware reinforcement learning framework with a hybrid GNN-Transformer mapper, boosting multi-accelerator inference throughput by 26.1% through applied AI/ML techniques and systems research.
- Extended a C++ compiler toolchain to deploy 30+ deep learning models (LLM/CV/MoE) as computation graphs on heterogeneous accelerators, improving deployment efficiency and execution accuracy.
- Reduced graph size by up to 90% via modularity-based clustering and co-authored STARDUST (IEEE CAL), cutting sample requirements by 15× and demonstrating improved evaluation efficiency for model behavior research.

Software Engineer Intern **| DHC Software** Beijing, CN | Jul. 2022 – Dec. 2022

- Designed a high-concurrency sales system to handle traffic spikes, boosting throughput by 100× (peak QPS from 500 to 50k+) by introducing Redis for hot-data caching and RocketMQ for asynchronous traffic leveling.
- Transitioned a monolithic backend into Java microservices (Spring Boot / Spring Cloud) with service discovery, API gateway routing, and fault-tolerant middleware, improving system resilience and scalability.
- Enhanced database performance through sharding and read–write separation, reducing latency under large-scale concurrent workloads.

Skills & Interests

Languages: Python, C++, Java, SQL, CUDA

ML/Systems: Reinforcement Learning, Optimization (MILP), GNN/Transformers, Compiler/Graph IR, Algorithms, Data Structures, Information Retrieval, Artificial Intelligence, Natural Language Processing,PyTorch, DeepSpeed,vLLM, AI/ML

Infra: Apache Airflow, Spark, Docker, Kubernetes, Redis, RocketMQ, Data Center, Distributed Computing, google cloud

Recognition: Kaggle Expert(with 2 silver medals in natural language processing and recommendation system)

Education

MS. Computer Engineering | George Washington University Washington,DC | 2023 - 2025