bauplan

Supporting Our AI Overlords (SAO)

Agents, data, and other stories.

The SAO team

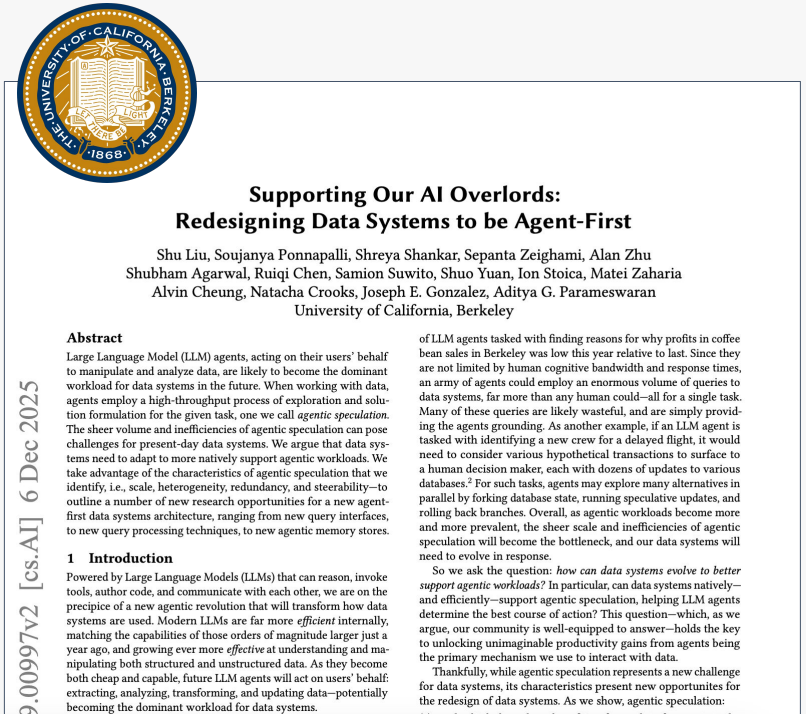
ACM CAIS, MAY 2026, San Jose, CA

The Times They Are A(gentic)-Changin'

Agents will be the primary **users** of data systems.

01 New abstractions

02 Unprecedented concurrency



Supporting Our AI Overlords: Redesigning Data Systems to be Agent-First

Shu Liu, Soujanya Ponnapalli, Shreya Shankar, Sepanta Zeighami, Alan Zhu
Shubham Agarwal, Ruiqi Chen, Samion Suwito, Shuo Yuan, Ion Stoica, Matei Zaharia
Alvin Cheung, Natacha Crooks, Joseph E. Gonzalez, Aditya G. Parameswaran
University of California, Berkeley

Abstract

Large Language Model (LLM) agents, acting on their users' behalf to manipulate and analyze data, are likely to become the dominant workload for data systems in the future. When working with data, agents employ a high-throughput process of exploration and solution formulation for the given task, one we call *agentic speculation*. The sheer volume and inefficiencies of agentic speculation can pose challenges for present-day data systems. We argue that data systems need to adapt to more natively support agentic workloads. We take advantage of the characteristics of agentic speculation that we identify, i.e., scale, heterogeneity, redundancy, and steerability—to outline a number of new research opportunities for a new agent-first data systems architecture, ranging from new query interfaces, to new query processing techniques, to new agentic memory stores.

1 Introduction

Powered by Large Language Models (LLMs) that can reason, invoke tools, author code, and communicate with each other, we are on the precipice of a new agentic revolution that will transform how data systems are used. Modern LLMs are far more *efficient* internally, matching the capabilities of those orders of magnitude larger just a year ago, and growing ever more *effective* at understanding and manipulating both structured and unstructured data. As they become both cheap and capable, future LLM agents will act on users' behalf: extracting, analyzing, transforming, and updating data—potentially becoming the dominant workload for data systems.

of LLM agents tasked with finding reasons for why profits in coffee bean sales in Berkeley was low this year relative to last. Since they are not limited by human cognitive bandwidth and response times, an army of agents could employ an enormous volume of queries to data systems, far more than any human could—all for a single task. Many of these queries are likely wasteful, and are simply providing the agents grounding. As another example, if an LLM agent is tasked with identifying a new crew for a delayed flight, it would need to consider various hypothetical transactions to surface to a human decision maker, each with dozens of updates to various databases.[2] For such tasks, agents may explore many alternatives in parallel by forking database state, running speculative updates, and rolling back branches. Overall, as agentic workloads become more and more prevalent, the sheer scale and inefficiencies of agentic speculation will become the bottleneck, and our data systems will need to evolve in response.

So we ask the question: *how can data systems evolve to better support agentic workloads?* In particular, can data systems natively—and efficiently—support agentic speculation, helping LLM agents determine the best course of action? This question—which, as we argue, our community is well-equipped to answer—holds the key to unlocking unimaginable productivity gains from agents being the primary mechanism we use to interact with data.

Thankfully, while agentic speculation represents a new challenge for data systems, its characteristics present new opportunites for the redesign of data systems. As we show, agentic speculation:

The Times They Are A(gentic)-Changin'

Agents will be the primary **builders** of data systems.

01 A vast design space

02 Most systems problems are verifiable

bauplan

together.ai

Stanford University

WISCONSIN

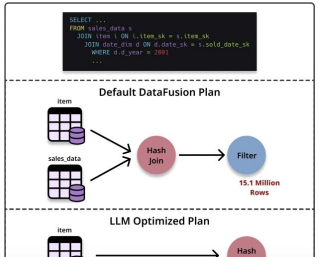
Making Databases Faster with LLM Evolutionary Sampling

Mehmet Hamza Erol [1] Xiangpeng Hao [2] Federico Bianchi [3] Ciro Greco [4] Jacopo Tagliabue [4] James Zou [3 1]

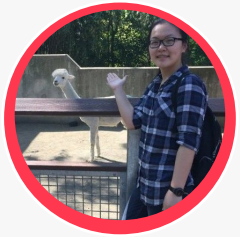
Abstract

Traditional query optimization relies on cost-based optimizers that estimate execution cost (e.g., runtime, memory, and I/O) using predefined heuristics and statistical models. Improving these heuristics requires substantial engineering effort, and even when implemented, these heuristics often cannot take into account semantic correlations in queries and schemas that could enable better physical plans. Using our DBPlanBench harness for the DataFusion engine, we expose the physical plan through a compact serialized representation and let the LLM propose localized edits that can be applied and executed. We then apply an evolutionary search over these edits to refine candidates across iterations. Our key insight is that LLMs can leverage semantic knowledge to identify and apply non-obvious optimizations, such as join orderings that minimize intermediate cardinalities. We obtain up to 4.78× speedups on some queries and we demonstrate a small-to-large workflow in which optimizations found on small databases transfer effectively to larger databases.

match user intent) and efficient (i.e., the query is processed as fast as possible). OLAP queries tend to be reused often (Van Renen et al., 2024; Tagliabue et al., 2024) and thus, there is real value in spending time to find optimized solutions. Even small improvements in planning would translate into massive efficiency gains at cloud scale, so it is not surprising that both industry and academia have devoted substantial resources to this task.



The SAO band



Elaine Ang

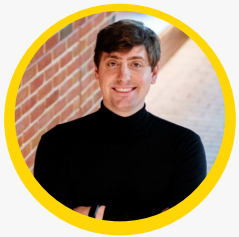
Columbia University

Shu Liu

UC Berkeley

Aditya Parameswaran

UC Berkeley



John Dickerson

Mozilla AI

Jonathan Frankle

Databricks

Jacopo Tagliabue

Bauplan Labs

Our guests: keynote speakers

Andy Pavlo

Carnegie Mellon University

Aaron Katz

ClickHouse

Nikita Shamgunov

databricks

Our guests: panelists

Ashish Kumar

MongoDB.

Anant Jhingran

IBM

Anupam Datta

snowflake

Junaid Ahmed

DATADOG

Ciro Greco

bauplan

Agenda

1:30 / 1:40 PM	Welcome
1:40 / 2:20 PM	Keynote: Aaron Katz
2:20 / 3:00 PM	Keynote: Andy Pavlo
3:00 / 3:30 PM	Break
3:30 / 4:30 PM	Spotlights and lightning talks
4:30 / 5:10 PM	Keynote: Nikita Shamgunov
5:10 / 6:00 PM	Panel
6:30 / 9:00 PM	Happy hour and awards

- Anam
- Austrian Supply Chain Intelligence Institute
- Bauplan Labs
- Columbia University
- Complexity Science Hub Vienna (CSH)
- Databricks
- Datadog
- Firebolt
- Google Cloud
- Hyperparam
- IBM
- LanceDB
- LithosAI, Inc.
- marimo
- Microsoft Research
- MongoDB
- Mozilla AI
- MotherDuck
- National University of Singapore
- NVIDIA
- Recce
- Redpanda
- Stanford University
- UIUC
- University of California, Irvine
- University of Edinburgh
- University of Illinois Urbana-Champaign
- University of Massachusetts Amherst
- University of New Hampshire
- University of Oxford

Tonight: happy hour

6:30 PM

📍 Dr. Funk

The SAO PCs

Program committee.

- **Aldrin Montana**
  Bauplan Labs
- **Alperen Keleş**
  University of Maryland
- **Bonnie Xu**
  OpenAI
- **Davide Eynard**
  Mozilla AI
- **Eugene Wu**
  Columbia University
- **Federico Bianchi**
  Together AI
- **Gaetano Rossiello**
  IBM
- **Joseph Axisa**
  Google
- **Nandana Mihindukulasooriya**
  IBM
- **Nicole Rose Schneider**
  University of Maryland
- **Sesh Nalla**
  Datadog
- **Stephanie Wang**
  MongoDB
- **Tao Ye**
  Lyft
- **Till Döhmen**
  MotherDuck

Our supporters

main sponsors

bauplan

Mozilla.ai

Best paper award

MongoDB.

Best student paper award

DATADOG

Want to chat?
Reach out.

jacopo.tagliabue@bauplanlabs.com