

Publish or Ship? The Research-to-Production Gap in Data Systems

Subarna Chatterjee*
IBM, WatsonX, USA
chatterjeesubarna@yahoo.com

PVLDB Reference Format:

Subarna Chatterjee. Publish or Ship? The Research-to-Production Gap in Data Systems. PVLDB, 19(12): XXX-XXX, 2026.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at URL_TO_YOUR_ARTIFACTS.

1 INTRODUCTION

Deepening Foundations vs. Chasing Scale: The Persistent Divide. The divide between how academia and industry view “large-scale systems” is not new, but unresolved [3, 4, 6]. While academia focuses on accelerating the science behind large-scale systems, the focus is on one algorithm at a time, mostly in a standalone manner. On the other hand, industry is all about scale in data and users and defining what technology looks like in practice. But the deliverable-driven pace leaves little or no room for the deep research to get that ultimate “elegant” solution. This panel focuses on *what each side is doing right, what they are getting wrong, and what it would actually take to build the bridge for the better.*

Multiple Realities, Same Gap. The panel is designed to include a diverse spectrum of perspectives [1, 2, 5] to address the *research-to-production gap*. For example, a researcher from academia focuses on monolithic systems due to limited funding to deploy a full-fledged serverless experimental setup. A pure industry practitioner has access to large-scale infrastructure but operates under time constraints, which leaves little room to explore or “be wrong” or ultimately land robust, lasting solutions. A startup founder with a Ph.D. is probably seeing the best of both worlds and struggling to bridge both worlds. A final-year PhD student is looking at the same divide but from a more personal angle, weighing the costs of each direction and where to commit. A hiring manager is still figuring out what a strong candidate actually looks like and what skills are missing on both sides.

Goal of the Panel. The goal is not to resolve the decade-old divide in a single panel session. It is to bring this divide to light, acknowledge the gap, and then engage in a candid discussion with genuine honesty about where we are and what we can do better. We believe that a well-structured public forum creates a shared space for researchers and practitioners to articulate this gap. At the end of the day, attendees should leave with a structured framing of this gap, a

few concrete directions to act on, and the motivation to carry that energy back into academic labs or industry environments and the broader community.

2 PANEL STRUCTURE

The panel is designed around a diverse set of panelists who can speak on behalf of their respective communities – an academic researcher, an industry practitioner, a bridge builder who has navigated both worlds, whether through an industrial research lab or direct experience on each side, a startup founder with a research background, and a final-year Ph.D. student or postdoctoral researcher who is at a transition point to the next career step.

Prompts for Opening Statements. To set the tone for a grounded discussion, the moderator will share the following questions with the panelists ahead of the session, prompting those to be included in their opening remarks.

- (1) What is your honest diagnosis of this gap?
- (2) What is your side genuinely doing right? What is the one thing you would critique on your side?
- (3) What do you think would actually work?

Moderator questions to get the discussion going. The moderator will structure the discussion in three phases. Phase 1 will start by establishing the problem space. Phase 2 will focus on deepening the discussion on the core issues, including direct exchange of questions between panelists. Finally, Phase 3 will conclude with reflections and actionable takeaways. Examples of moderator questions are:

- (1) How realistic are academic benchmarks, e.g., TPC-H, YCSB? Do they truly reflect the complexity of real-world workloads? If not, what would it take to address this?
- (2) Researchers value novelty while engineers get promoted for shipping. But neither gets rewarded for closing the gap. So why would anything change, and what would have to change first?
- (3) Knowledge is being generated at a production scale every day that no research environment can fully replicate, but most of it does not reach the broader community of students or researchers. What are the real constraints here? How should the balance exist between openness and legitimate business logic?

3 PREFERRED PANELISTS

The set of panelists is as follows. The final participation of the panelists is subject to availability and confirmation.

- (1) Academic Researcher: Natacha Crooks, UC Berkeley
- (2) Academic Researcher: Ryan Marcus, University of Pennsylvania
- (3) Industry Practitioner: Sudipto Das, Databricks
- (4) Research Lab: Yuanyuan Tian, Microsoft Research

*This panel proposal reflects the author’s independent research interests. The institutional affiliation is listed for identification purposes only.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 19, No. 12 ISSN 2150-8097.
doi:XX.XX/XXX.XX

- (5) **Industry Practitioner or Startup Founder:** To be confirmed

4 MODERATOR

Subarna Chatterjee is a research software engineer at IBM WatsonX (Datastax), where she works at the intersection of large-scale data systems and generative AI infrastructure. Currently, Subarna designs cloud-native, multi-tenant database architectures through serverless computing to support big data and AI workloads. Before switching gears into industry, Subarna spent four years as a Postdoctoral Fellow at the Data Systems Laboratory at Harvard University and a year as a Postdoctoral Researcher at Inria Rennes in France. She holds a PhD in Computer Science and Engineering from the Indian Institute of Technology Kharagpur. Her post-doc research spans self-designing storage engines, key-value store optimization, range filters for LSM-based systems, and data stream processing for IoT applications. Having moved from academia to

industry research, Subarna brings a lived perspective on what this gap looks like and becomes a natural voice in a conversation about how these two worlds can collaborate.

REFERENCES

- [1] 2020. Research in Industry or Academia? <https://tanyajonker.com/essays/research-industry-academia>.
- [2] 2023. Study: Industry now dominates AI research. <https://mitsloan.mit.edu/ideas-made-to-matter/study-industry-now-dominates-ai-research>.
- [3] 2024. AI's computing gap: academics lack access to powerful chips needed for research. <https://www.nature.com/articles/d41586-024-03792-6>.
- [4] 2024. Bridging the Divide: Academia vs. Industry in the Age of Accelerated Innovation. <https://www.linkedin.com/pulse/bridging-divide-academia-vs-industry-age-accelerated-innovation-phd-blkuc/>.
- [5] 2025. Bridging the Gap Between Industry and Academia in Computer Science. https://medium.com/@ahmedelhadi_16675/bridging-the-gap-between-industry-and-academia-in-computer-science-ca268e1d1070.
- [6] 2025. Bridging the Gap: Translating Academic Research into Industry Innovations. <https://www.linkedin.com/pulse/bridging-divide-academia-vs-industry-age-accelerated-innovation-phd-blkuc/>.