FSE 2026 Workshop: Data Intensive Software Development (DISD)

ANONYMOUS AUTHOR(S)

ACM Reference Format:
Anonymous Author(s). 2025. FSE 2026 Workshop: Data Intensive Software Development (DISD). 1, 1 (October 2025), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Modern software systems are becoming increasingly data-intensive. They process massive quantities of data and also produce large-scale, machine-generated data, such as traces and telemetry, which is used for runtime monitoring, log analysis, debugging, and testing. The analytical pipelines are also performance-critical, and their behavior is highly dependent on the scale and content of the data they process.

Due to this massive scale, it is essential to incorporate intelligent AI agents that can not only process the data but also reason about it within the context of the software. Integrating these AI-driven, data-informed feedback loops is the key to enabling autonomous software operations and providing context-aware guidance to developers. However, achieving trustworthy AI-native software engineering requires addressing critical challenges that directly impact the robustness, reliability, and trustworthiness of data-intensive systems

- Data-dependent uncertainty: The dynamic nature of large-scale, machine-generated data makes it difficult to design comprehensive test cases and debug unexpected behaviors. Further, the use of AI and agentic approaches, both (1) in these software systems and (2) for debugging and monitoring large-scale data for detecting anomalies, can amplify the uncertainty in diagnosis.
- Massive scale: The massive volume of data overwhelms traditional methods, making analysis computationally infeasible. This requires systems-level optimization for storage, tracing, and monitoring as critical system behaviors only manifest at production scale.
- Stringent latency requirement: The real-time or near real-time requirements of data-intensive applications make traditional offline analysis too slow to detect and respond to issues timely. The long running time of data-intensive applications makes techniques such as fuzzing not easily applicable at scale.
- Robustness under evolving conditions: Data-intensive systems operate in dynamic environments where both data distributions and infrastructure conditions shift unpredictably. Trustworthy AI approaches must remain stable and reliable despite distribution drift, adversarial inputs, and noisy runtime data.

ACM XXXX-XXXX/2025/10-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Workshop Information

This workshop community will provide a forum for researchers and practitioners to advance software engineering research around data-intensive software development. We will invite discussions that explore fundamental ideas, practical solutions, and cross-domain innovations in this space, including but not limited to:

- Data-intensive software testing, debugging, runtime monitoring, and log analytics
- Semantics lifting from systems-generated data
- Modelling application behavior via data-system coordination
- The use of AI, ML, and agentic approaches for monitoring, debugging, and testing systems-generated data
- Algorithms and foundations for testing and debugging performance-critical systems
- Benchmarks for testing, debugging, and analysis of systems-generated data

Goal: These research topics are vital for trustworthy AI incorporation in data-intensive software development. Due to data sensitivity, agent systems' interaction logs, large-scale logs, input data, queries, workloads for big data analytics, and associated bugs, are often not released to the public. To drive **trustworthy AI and agentic AI incorporation in data intensive software development**, we are in the urgent need of forming a research community (a consortium) that **curates and maintains industry-scale, synthetic datasets and benchmarks** .

Organizers.

- Lionel Briand (lbriand@uottawa.ca), Lero Centre, Ireland
- Tse-Hsun (Peter) Chen (peterc@encs.concordia.ca), Concordia University, Canada
- Muhammad Ali Gulzar, (gulzar@cs.vt.edu), Virginia Tech
- Yintong Huo (ythuo@smu.edu.sg), Singapore Management University
- Miryung Kim (miryung@cs.ucla.edu), UCLA
- Michael Lyu (lyu@cse.cuhk.edu.hk), Chinese University of Hong Kong
- Weiyi Shang (wshang@uwaterloo.ca), U. Waterloo

Format and required services. The proposed workshop will be structured as a 1-day event designed to maximize participation, knowledge exchange, and collaborative problem-solving around data-intensive software. The program will feature paper presentations in the nature of research idea proposals or position papers), along with interactive group activities designed to stimulate discussion and idea generation.

Keynote session (morning 1). The morning session of the workshop will begin with a keynote, setting the stage by outlining the major challenges and opportunities in data-intensive software engineering. This shared foundation will help align participants around the central themes of the workshop while sparking ideas for further discussion. We plan to invite keynote speakers who can discuss the scale of AI incorporation in data intensive software development in industry such as Tim Kraska (Amazon Web Services), Dongmei Zhang (Microsoft), Nachiappan Nagappan (Meta), Chao Peng (ByteDance), Ahmed Hassan (Huawei), Jeromy Carriere (DataDog).

Paper presentation session (morning 2). Following the keynote, in the second session in the morning participants will present their papers. We will be soliciting papers in the type of research idea proposals or position papers to stimulate discussions. These presentations will provide concrete starting points for the collaborative activities scheduled later in the day.

By the end of the morning session, the organizers will finalize the themes that emerged from the keynote and proposal presentations. The afternoon breakout groups will be structured around well-defined themes of interest.

Breakout session (afternoon 1). Based on the consolidated four to five themes created by the end of the morning session, and attendees will divide into smaller breakout rooms. Within each group, participants will explore shared challenges, identify synergies, and generate new ideas that may not surface in larger plenary discussions. At the end of this breakout session, our goal is to form a team to volunteer to write a section in a community driven article on **challenges and reflections in data-intensive software development.**

Round-table session (afternoon 2). After the breakout discussions, all groups will reconvene in the main session to share key insights and perspectives across themes. Each group will briefly present their key findings and emerging ideas. The round table format will encourage an open, interactive exchange, highlight connections between themes, and debate differing viewpoints. The expected deliverable for this round-table session is **a cohesive outline for a community-driven article** on challenges and reflections in data-intensive SW development.

Publicity Plans. Workshop publicity will span multiple coordinated channels. Announcements will be distributed via SIGSOFT's SEWORLD mailing list to reach the broader software engineering community. We will regularly post updates on X and LinkedIn to highlight key dates for the workshops and feature our keynote speakers. Additionally, we will promote the workshops through FSE's official channels, including the conference website and social media platforms. We will also consider sending direct invitations to FSE 2026 authors across various tracks to encourage their participation, submissions, and attendance at the workshop.

Publications The workshop organizers will not be submitting papers to their own workshop. After the workshop, the organizers will summarize advances and challenges in the area of data-intensive software development as a recurring **community-driven article** that forms the body of knowledge in this area. We have received a positive response from the editor of chief at IEEE Software on disseminating the workshop outcome report, as a recurring featured article on "Challenges and reflections on data intensive SW development."

Statement of Overlap With Prior Workshops This workshop is different from a continuing 20 years history of Mining Software Repositories[1] community by focusing on systems-generated data. The AIOps workshop[2] focuses on applying AI to cloud operations, while our workshop is not limited to cloud operations but building, debugging, and testing data-intensive SW development and AI incorporation in such context. Also our goal is to create a community to curate and maintain industry-scale synthetic dataset and benchmarks to drive research in AI incorporated, data-intensive SW development. Unlike the Workshop on Distributed Software Development, Software Ecosystems and Systems-of-Systems[3], which emphasizes distributed development and ecosystems, this workshop target the complexities of large-scale data processing. Compared to the International Workshop on Agentic Engineering[4], which is centered on autonomous agents, our scope is broader and data-driven. While LLM4Code[5] is focused on large language models for code-centric tasks, this workshop address challenges in building, debugging and testing data-intensive pipelines.

3 Paper Selection Procedure

The workshop welcomes the following submission types:

- Position Statements (1 to 2 pages): early-stage ideas, industrial perspectives, proposals to contribute to the working group report.

[1] https://www.msrconf.org
[2] https://cloudintelligenceworkshop.org
[3] https://conf.researchr.org/home/sesos-wdes-2021
[4] https://conf.researchr.org/home/icse-2026/agent-2026
[5] https://llm4code.github.io

- Short Papers (up to 5 pages): early-stage ideas, visions, or tool reports.
- Full Papers (6–10 pages): novel approaches, frameworks, or evaluations.
- arXiv Presentations (non-archival): recent preprints for open discussion.

Authors of accepted position statements will collaborate to combine and expand their experience and insight into a community paper to be submitted to IEEE Software or a similar venue. The goal of this paper is to provide a shared perspective on the current state of the field and offer a roadmap for the future.

All papers will be submitted via HotCRP and will be reviewed in a double-blind manner. The submission should comply with the ACM format (in line with FSE 2026) and should present the original contribution. At least one author of each accepted paper must register for the workshop and present the paper there.

All accepted papers, except for the position statements and arXiv presentation, will appear in the FSE 2026 workshop proceedings by default. For the non-archival papers and position statements, the camera-ready version will only be posted/advertised on our workshop website. Please note that no matter which option you choose (archival or non-archival), the submission should be fully original (not accepted/published anywhere else) by the submission time, and at least one author has to register for the workshop and present. The official publication date of the workshop proceedings is the date the proceedings are made available by ACM.

TODO: Expected attendees

TODO: Format: 1 day?

TODO: PC formation and Publicity: Ian + Yintong + Gulzar, merge with keynote list

TODO: merge the two lists Domenico Bianculli (University of Luxembourg),
Michael Pradel (CISPA),
David Lo (Singapore Management University),
Xiaofei Xie (Singapore Management University),
Lingming Zhang (UIUC),
Chao Peng (Bytedance),
Ying Li (Peking University),
Jie Zhang (KCL),
Pinjia He (Chinese University of Hong Kong, Shenzhen)
Shan Lu (University of Chicago)
Heng Li (Polytechnique Montreal)
Junwen Yang (Meta)
Jeromy Carriere (https://www.linkedin.com/in/jeromycarriere/)

TODO: miryung:we probably need some more industry participants? more european participants?

4 Working Group Titles for "What's going on in data intensive SW development" curation paper

4.1 Working Group 1: Data-Intensive Benchmarks, Bugs, and Oracles

Data-intensive software systems disproportionately lack benchmarks, bug repositories, and domain-specific oracles. For example, widely used frameworks such as Apache Hadoop and Spark are still evaluated against decades-old SQL benchmarks (e.g., TPC) or unrealistically simple programs (e.g., sort). At the same time, we have little systematic knowledge of the bugs that occur in these frameworks and the applications built atop them, nor of the oracles typically used to detect such bugs. As a result, software engineering research in this domain has remained underrepresented. A core objective of this workshop is to stimulate open discussion on pathways to design natural, real-world-inspired, and scalable benchmark programs for data-intensive stacks, their bugs, and their oracles, with the aim of advancing *correctness*, *performance*, and *scalability* testing. A key element for this working group is to investigate synergistic approaches to acquire such benchmarks from commercial and industry stakeholders (e.g., Amazon EMR, Google Dataflow, Databricks, Snowflake) while respecting intellectual property, data privacy, and business confidentiality constraints.

TODO: homework: title and one sentence description

TODO: Data-intensive Benchmarks Gulzar Miryung, bugs oracles,

TODO: Synergy with AI/ Agentic Peter

TODO: AI-Ops, runtime monitoring, Yintong

4.2 Working Group2: Data-Intensive Software Runtime Monitoring

Modern software systems generate vast amounts of runtime data, but monitoring often treats systems as a monolith, obscuring the component interactions necessary for precise failure attribution. This challenge is further amplified by the integration of recent black-box AI/LLM plugins, which complicates root cause analysis. This working group will tackle these challenges by exploring methods to *instrument and exploit runtime data for monitoring*. We will focus on creating benchmarks and analyzing techniques for fine-grained failure attribution and localization, especially within complex systems that combine traditional software with AI components.

4.3 Working Group X: data-intensive systems debugging and testing

Debugging and testing data-intensive systems present unique challenges that set them apart from traditional software. The scale, heterogeneity, and continuous flow of data make it difficult to reproduce failures, isolate root causes, or ensure comprehensive test coverage. Data quality itself can become a hidden source of errors, where issues may arise not from the code but from anomalies, inconsistencies, or biases in massive datasets. Moreover, the interplay between distributed components and complex data pipelines creates emergent behaviors that are hard to anticipate with conventional testing techniques. TODO: I wonder wehther we should combine 4.1 and 4.3

4.4 Working Group X: Non-functional quality attributes of data-intensive systems

Non-functional aspects of data-intensive systems, such as security, privacy, performance, scalability, reliability, and energy efficiency, are central concerns, as they directly shape the trustworthiness and usability of these systems. Managing massive and heterogeneous data at scale requires not only functional correctness but also careful attention to how systems safeguard sensitive information, maintain resilience under heavy loads, and deliver predictable performance in dynamic environments.

4.5 Working Group X: AI-Agentic Systems as Data Producers and Consumers

TODO: to be changedThe rise of AI-agentic systems creates a unique dual role: they are both sophisticated consumers of data-intensive infrastructure and significant generators of complex behavioral data. This working group will explore the complete lifecycle of data in AI-agentic environments, where the interactions, decisions, and reasoning traces of AI agents themselves become valuable telemetry for system understanding. We will address challenges such as instrumenting multi-agent workflows, capturing meaningful agent decision logs, and leveraging this AI-generated data to improve system reliability, performance, and transparency. Key focus areas include developing standards for agent telemetry, creating benchmarks that include agent interaction patterns, and designing monitoring solutions that can attribute system behavior to specific agent decisions or emergent collective behaviors. This group will bridge the gap between traditional software monitoring and the new paradigms required for autonomous, AI-driven systems. TODO: keep the title only and move the merged content or content on the website TODO: Gulzar:website TODO:

Ian: differentiation from existing work TODO: Yintong, Peter, Me taking a pass in the oder, will have a final version before Oct 3 meeting