SUBHASIS DASGUPTA, Ph.D.

DATA PLATFORM & DATABASE SYSTEMS ENGINEER | STARTUP BUILDER | QUERY PERFORMANCE

San Diego, CA • +1 858-366-3393 • sudasgupta@ucsd.edu

Database systems and data-platform engineer with 20+ years of experience building distributed data, cloud, and research platforms from early-stage concepts through sustained operation. Deep expertise in query processing, ingestion, heterogeneous databases, search, knowledge graphs, streaming data, performance engineering, and platform observability. At Kaavo, served as the first employee and later Founding Director of the India operation, helping build an early application-centric multi-cloud management product. At UC San Diego and SDSC, leads architecture and implementation of production-oriented scientific data platforms, co-invented patented polystore ingestion and query-processing technologies, and works hands-on across Python, C++, Java, SQL, PostgreSQL, ClickHouse, Kubernetes, and distributed workflows.

CORE EXPERTISE

Database Internals & Query Processing: PostgreSQL; ClickHouse; DuckDB; Neo4j; query planning; indexing; operators; polystores; SQL optimization

Data Platform Engineering: streaming ingestion; ETL/ELT; schema evolution; metadata; APIs; event pipelines; data quality; analytical serving

Distributed Systems & Cloud: Kubernetes; Docker; Linux; distributed computing; public/private/hybrid cloud; service orchestration; deployment automation

Observability & Reliability: Prometheus; Grafana; Fluent Bit; instrumentation; anomaly detection; health monitoring; bottleneck and incident diagnosis

Programming & Systems: Python; C++; Java; SQL; PL/Python; Perl; parsers; APIs; automation and testing frameworks

Search, Graph & AI Data: Solr; Weaviate; Neo4j; semantic retrieval; knowledge graphs; ontology grounding; model-data integration

Workflow & Analytics: Nextflow; Snakemake; reproducible pipelines; federated access; large-scale scientific and telemetry processing

Security & Governance: Keycloak; Trivy; GPG; policy-aware sharing; secure containers; ontology-based authorization; data governance

PROFESSIONAL EXPERIENCE

San Diego Supercomputer Center & University of California San Diego | La Jolla, CA **2015–Present**

Computational and Data Science Research Specialist IV (2017–Present); Assistant Project Scientist (2019–Present); Postdoctoral Researcher (2015–2017)

- Architect and implement distributed data platforms across SDSC and UC San Diego, integrating ingestion pipelines, databases, search systems, knowledge graphs, APIs, containers, workflow engines, and analytical services.
- Design and troubleshoot high-volume data workflows, tracing latency, correctness, and reliability issues across application, pipeline, database, storage, container, and compute layers.
- Build production-oriented Kubernetes and Docker services with Prometheus/Grafana/Fluent Bit observability, automated deployment, security scanning, identity management, and CI/CD workflows.
- Lead database and pipeline performance engineering for PostgreSQL, ClickHouse, DuckDB, graph systems, and telemetry workloads; analyze query plans, code paths, memory, I/O, throughput, latency, and scaling behavior.
- Architected the National Data Platform federated-search capability, linking distributed metadata catalogs, text and semantic search, APIs, schema normalization, and policy-aware access.
- Led TemPredict platform architecture for continuous event and wearable-sensor ingestion, harmonization, monitoring, data-quality control, and analytical workflows supporting multi-institution research.
- Build data and retrieval services for AI and agent workflows, integrating open-source models with databases, knowledge graphs, scientific APIs, evaluation pipelines, and reproducible execution.
- Lead cross-functional architecture and implementation with engineers, scientists, and external partners; mentor developers and translate ambiguous requirements into reusable data products and platforms.

Kaavo Inc. / Kaavo Systems India Pvt. Ltd. | United States & India |
Founding-stage cloud startup **2008–2012**

First Employee; Senior Systems Engineer and Founding Director, India

- Joined Kaavo as its first employee and helped build the company from the founding stage, contributing hands-on systems engineering, product development, deployment automation, and early operational execution.
- Co-developed IMOD, an early application-centric cloud-management platform that automated deployment and runtime management of multi-tier applications across public, private, and hybrid clouds; worked across architecture, provisioning, configuration, monitoring, releases, and customer troubleshooting.
- Established and led the India engineering operation as Founding Director, combining hands-on development with recruiting, team formation, delivery coordination, and startup execution in a resource-constrained environment.

Indian Statistical Institute | Kolkata, India **2012–2015**

Project-Linked Research Fellow

- Researched ontology-aware query processing and fine-grained access control for large digital-library metadata and graph structures; developed prototypes, formal models, and peer-reviewed publications.

SELECTED DATA PLATFORM AND DATABASE CONTRIBUTIONS

High-Volume Data and Event Workloads. Design ingestion, transformation, and analytical workflows for telemetry, wearable, scientific, graph, and text data; diagnose latency, schema, resource, and scaling bottlenecks across the stack.

Data Platform Operations. Design ingestion, search, transformation, workflow, monitoring, identity, security, and API services; support reproducible Kubernetes/Docker deployments with CI/CD, instrumentation, and operational telemetry.

AWESOME Polystore. Co-invented ingestion and query-processing techniques across databases, graph stores, text indexes, and analytical engines; built planning and execution capabilities for scientific applications.

National Data Platform Search. Architected federated search across distributed scientific data and APIs using metadata normalization, text and semantic retrieval, policy-aware access, and reusable discovery services.

TemPredict. Led architecture for continuous wearable-data ingestion, harmonization, monitoring, and analysis, enabling longitudinal research and predictive-health model development.

Quantum Data Hub. Designed collaborative data infrastructure for quantum-materials research, supporting metadata federation, discovery, visualization, and reproducible analysis.

AI-Ready Data Services. Develop retrieval, ontology-grounded extraction, knowledge-graph, evaluation, and multi-step workflow pipelines that connect models with governed data and APIs.

Database Performance Engineering. Design benchmarks and instrumentation for PostgreSQL, ClickHouse, DuckDB, graph, and telemetry systems; analyze plans, operators, code paths, throughput, latency, memory, and I/O scaling.

EARLIER INDUSTRY AND ACADEMIC EXPERIENCE

DataInfoCom Software Solutions — Software Engineer (R&D). Built ETL and Java APIs for ARIMA-based prediction in a business-intelligence product.	2007–2008
Marine Engineering and Research Institute, Indian Maritime University — Assistant Professor, Computer Science & Engineering.	2006–2007
Connectiva Systems — Technical Specialist.	2006
Centre for Mobile Computing and Communication, Jadavpur University — Research Engineer in grid and mobile computing.	2004–2006
Rady School of Management, UC San Diego — Lecturer.	2019

LEADERSHIP, COLLABORATION, AND SERVICE

- Mentor engineers and students in database internals, query processing, cloud automation, data pipelines, platform operations, and applied AI; lead technical coordination, code review, and architecture discussions.
- Program committee member for ICDE and other computing conferences; reviewer for leading IEEE distributed-systems, security, and computing journals.
- Collaborate across computer science, medicine, materials science, social science, and cyberinfrastructure; communicate technical tradeoffs to scientific and sponsor audiences.

PATENTS

- Data Ingestion into a Polystore — U.S. patent application US201762594408P.

- Query Processing in a Polystore — U.S. patent application US20220083552P.

SELECTED PUBLICATIONS

- X. Zheng, S. Dasgupta, and A. Gupta, “P2KG: Declarative Construction and Quality Evaluation of Knowledge Graphs from Polystores,” ADBIS, 2023.
- S. Purawat et al., “TemPredict: A Big Data Analytical Platform for Scalable Exploration and Monitoring of Personalized Multimodal Data for COVID-19,” IEEE International Conference on Big Data, 2021.
- S. Purawat et al., “Quantum Data Hub: A Collaborative Data and Analysis Platform for Quantum Material Science,” International Conference on Computational Science, 2021.
- S. Dasgupta, A. Bagchi, and A. Gupta, “Ingesting High-Velocity Streaming Graphs from Social Media Sources,” IEEE eScience, 2019.
- S. Dasgupta, C. McKay, and A. Gupta, “Generating Polystore Ingestion Plans: A Demonstration with the AWESOME System,” IEEE International Conference on Big Data, 2017.
- S. Dasgupta, K. Coakley, and A. Gupta, “Analytics-Driven Data Ingestion and Derivation in the AWESOME Polystore,” IEEE International Conference on Big Data, 2016.
- A. Gupta, S. Dasgupta, and A. Bagchi, “PROFORMA: Proactive Forensics with Message Analytics,” IEEE Security & Privacy 15(6), 33–41, 2017.
- R. K. Maji et al., “PVT: An Efficient Computational Procedure to Speed Up Next-Generation Sequence Analysis,” BMC Bioinformatics 15, 167, 2014.
- S. Khatua, S. Dasgupta, and N. Mukherjee, “Pervasive Access to the Data Grid,” International Conference on Grid Computing Applications, 2006.
- S. Dasgupta and A. Bagchi, “Controlling Access to a Digital Library Ontology: A Graph Transformation Approach,” International Journal of Next-Generation Computing 5(1), 2014.
- S. Dasgupta and A. Gupta, “Discovering Interesting Subgraphs in Social Media Networks,” ASONAM, 2020.
- J. Gao, S. Dasgupta, and A. Gupta, “Multi-Model Investigative Exploration of Social Media Data with BOUTIQUE,” IEEE eScience, 2019.

SELECTED ONGOING SYSTEMS WORK

- Cost-Based Planning for Natural-Language Queries with Open-World Predicates — designing a planner that combines relational operators, external semantic operators, cost estimation, and plan evaluation.
- Database Benchmarking and Instrumentation — comparing PostgreSQL, ClickHouse, DuckDB, and related engines using query plans, code instrumentation, system telemetry, and CPU/GPU performance profiles.
- Agentic Data Infrastructure — designing layered agent architectures, distributed execution services, resource connectors, and database/knowledge services for long-running workflows.

EDUCATION

Ph.D. in Engineering, Jadavpur University — Dissertation: “On the Design of an Ontology-Based Access Control Model: A Digital Library Perspective.” Advisors: Aditya Bagchi and Chandan Mazumdar. **2016**

M.E. in Computer Science & Engineering, Jadavpur University — Dissertation: “Performance-Based Scheduling in a Grid Environment.” **2005**

B.E. in Computer Science & Engineering, Asansol Engineering College, Burdwan University. **2003**