Subhadip Mitra

Engineering Leader & AI Systems Architect

Singapore | contact@subhadipmitra.com
LinkedIn: linkedin.com/in/subhadip-mitra | GitHub: github.com/bassrehab

Last updated: December 25, 2025

Professional Summary

Senior Engineering Leader with 15+ years of experience bridging fundamental AI research and enterprise-scale system delivery. Currently leading Google Cloud's Data & Analytics practice for Southeast Asia while driving internal innovations on LLM inference efficiency and multi-agent systems. Proven track record of operating as a "Player-Coach": managing regional engineering portfolios while simultaneously architecting and patenting novel frameworks (UPIR, ARTEMIS, FTCS, Speculative Decoding).

Professional Experience

Google Cloud - Professional Services Organization: Head of Data & Analytics, Southeast Asia — Site Lead, PSO Southeast Asia *January 2021 – Present*

Dual-track role combining technical innovation leadership with regional delivery management. Built Google Cloud's Data Analytics practice across Southeast Asia while serving as Site Lead overseeing cross-practice operations. Member of delta - Google Cloud's innovation and transformation team architecting enterprise AI solutions at scale.

Strategic Leadership & Delivery:

- Built Data Analytics practice for Southeast Asia from 0 to 1, recruiting and developing engineering talent while establishing the region's premier capability serving strategic enterprise clients across 6 countries.
- Serve as Site Lead overseeing delivery governance across all 7 PSO practices (Data Analytics, AI/ML, Infrastructure, Security, Enterprise Architecture, Application Development, Delivery Management) in Southeast Asia, owning utilization and CSAT metrics (97%) while driving strategic pursuits and contributing to 100% annual revenue target attainment.
- Direct $XXM+ Data Analytics delivery portfolio across JAPAC while simultaneously overseeing $XXM+ cross-practice portfolio as regional Site Lead.
- Led critical engagements for JAPAC strategic accounts including major financial services institutions, telcos and consumer electronics manufacturers, ensuring delivery excellence and client success.
- Partner with Sales leadership on strategic pursuits and collaborate with Product Engineering to shape platform roadmap based on field insights and customer requirements.
- Spearheaded cross-practice rescue operations for at-risk enterprise accounts with multi-million dollar project values, recovering strategic customers and converting potential platform exits into long-term partnerships.
- Delivered first-of-kind solutions including GenAI-powered reconciliation framework for a major airline (now replicated across JAPAC), large-scale ML platform migrations (30K+ notebooks), and petabyte-scale data platform modernizations.
- Partner with C-level stakeholders (CTOs, CDOs) to define data modernization and AI transformation roadmaps, translating technical capabilities into business outcomes.
- Pioneered agentic AI adoption across all 7 PSO practices (Data Analytics, AI/ML, Infrastructure, Security, Enterprise Architecture, Application Development, Delivery Management) and 6 JAPAC sub-regions, building SDKs, agent catalog, and standardized templates while designing reusable governance frameworks that accelerated innovation and reduced delivery costs.
- Built agentic tool suites including architecture discovery (100M+ node graph modeling), automated data pipeline generation, and platform cleanup agents that recovered multi-million dollar at-risk engagements and secured significant long-term cloud commitments.

- Built Data Strategy competency from 0, delivering 8-figure pursuit value across 14 strategic pitches in Asia Pacific while establishing critical data assets and new GTM offerings.

Technical Innovation & Research:

- 5 Google Technical Disclosures on AI and distributed systems - UPIR (automated system synthesis, 274x speedup), FTCS (context architecture for AI agents), ARTEMIS (multi-agent debate framework), ETLC (data processing for GenAI), and LLM inference optimization (speculative decoding, custom Triton kernels).
- Industry-agnostic agentic AI for enterprise trust decisions. APLS self-learning + cascade routing achieving 86% cost reduction, sub-50ms latency. Won Google Cloud PSO Hackathon JAPAC, qualified for World Finals.

Standard Chartered Bank: Principal Engineer - Data & Analytics Transformation *January 2019 – January 2021*

Led enterprise-wide AI and data platform development serving 11 markets and 1200+ global users, delivering technical excellence while influencing C-suite data strategy.

- Delivered a Self-Service ML Platform that reduced model development time from 6 months to 1 week
- Designed credit risk AI models integrating alternative data sources, improving accuracy by 15%
- Modernized MarTech infrastructure, driving 30% increase in customer acquisition

Think Big Analytics (a Teradata company): Principal Data Engineer / Solution Architect *January 2017 – January 2019*

Architected enterprise-scale data solutions for Fortune 500 clients across APAC, designing scalable platforms with measurable business impact.

- Engineered 5 high-performance data lakes processing 1.2 PB/hour, achieving 20% optimization
- Built real-time fraud detection systems, reducing false positives by 60% and saving $XM annually
- Designed enterprise architectures supporting global Fortune 500 clients across APAC

Various Companies: Software Engineering, Architecture and Technical Consulting Roles *January 2010 – January 2017*

Progressively advanced through roles in software development, systems integration, and technical consulting within financial services and algorithmic trading domains.

Research & Open Source Engineering

Spark LLM Eval - Distributed Evaluation Framework
Distributed LLM evaluation framework built on Apache Spark for enterprise-scale model assessment. Addresses the gap in evaluating LLMs at scale with statistical rigor, integrating seamlessly with Databricks infrastructure.
GitHub: https://github.com/bassrehab/spark-llm-eval | /blog/2025/building-spark-llm-eval

CatchMe - Intelligent Trust Engine
First of a kind, industry agnostic hybrid agentic AI decisioning system across Finance, Healthcare, Insurance, Cybersecurity, and Supply Chain. Uses adversarial debate protocols (prosecutor/defense/judge) to qualify events/anomalies and build audit trails for regulated environments.

Google Technical Disclosures - Pending (APLS & Cascade Routing)

LLM Inference Efficiency Research
Research implementations addressing the fundamental bottleneck in LLM inference: memory-bandwidth constraints rather than compute limits. Explores acceleration through speculative decoding, custom GPU kernels, and quantization strategies.

GitHub: https://github.com/bassrehab/speculative-decoding | GitHub: https://github.com/bassrehab/triton-kernels
Google Technical Disclosure - Pending

AI Metacognition Toolkit
Activation-level detection of sandbagging, deception, and situational awareness in LLMs. Linear probes achieve 90-96% accuracy across Mistral, Gemma, and Qwen models. Includes steering vectors for runtime behavior control.
PyPI: https://pypi.org/project/ai-metacognition-toolkit/ | GitHub: https://github.com/bassrehab/ai-metacognition-toolkit | Docs: https://ai-metacognition-toolkit.subhadipmitra.com/ | /blog/2025/detecting-ai-sandbagging/

Steering Vectors for Agent Behavior Control
Runtime control of LLM agent behaviors through activation steering vectors - modifying model outputs at inference time without retraining. Demonstrates more calibrated control than traditional prompting approaches with LangChain integration.
GitHub: https://github.com/bassrehab/steering-vectors-agents | /blog/2025/steering-vectors-agents

Publications & Technical Disclosures

UPIR: Automated Synthesis and Verification of Distributed Systems
Google, Technical Disclosure Commons, November 2025
https://www.tdcommons.org/dpubs_series/8852/

ETLC: A Context-First Approach to Data Processing in the Generative AI Era
Google Cloud, May 2025
https://services.google.com/fh/files/blogs/etlc_full_paper.pdf

Field-Theoretic Context System (FTCS)
Google, Technical Disclosure Commons, May 2025
https://www.tdcommons.org/dpubs_series/8022/

ARTEMIS - Adaptive Multi-agent Debate Framework
Google, Technical Disclosure Commons, January 2025
https://www.tdcommons.org/dpubs_series/7729/

Data Monetization Strategy for Enterprises
BITS Pilani, December 2023
https://www.researchgate.net/publication/376557741_Data_Monetization_Strategy_for_Enterprises

OConsent: Open Consent Protocol for Privacy and Consent Management with Blockchain
BITS Pilani, December 2021
https://arxiv.org/abs/2201.01326

Education

MBA, Business Analytics — Birla Institute of Technology and Science, Pilani
MTech, Software Systems — Birla Institute of Technology and Science, Pilani

Technical Skills

Technology Leadership & Strategy: Enterprise Architecture, Digital Transformation, AI & Data Strategy, C-Suite Advisory, Innovation Leadership, Strategic Planning

Data Engineering & Architecture: Data Pipelines, Real-Time Processing, Data Mesh & Fabric, Data Governance, Apache Spark, Delta Lake, Apache Kafka, Apache Iceberg

Generative AI & Machine Learning: Multi-Agent Systems, Large Language Models, RAG Architecture, Vector Databases, PyTorch, LangChain, LangGraph, LlamaIndex, Google ADK, MCP, A2A Protocol, MLflow, LLMOps

Cloud Platforms & Infrastructure: Google Cloud Platform, BigQuery, Vertex AI, Dataproc, Cloud Composer, GKE, Terraform, Kubernetes

Programming & Development: Python, SQL, Scala, Triton, CUDA, Algorithm Design, Formal Verification, Program Synthesis, Distributed Systems