Saswata Dutta

✉ saswat.dutta@gmail.com   🔗 saswata-dutta   📞 +91-8178349520

SKILLS

- Large Scale Data Intensive Applications, Big Data, Database Scaling
- Machine Learning, Machine Learning Platform, MLOps
- NLP, Generative AI, LLM, RAG, LLMOps, Chatbots
- AWS, GCP, Azure
- Leading technical charter for teams with 30+ engineers.

EXPERIENCE

Coinbase

Staff Software Engineer   *Oct 2023 - Present*

- Built a **scalable, multi-cloud LLM platform** , enabling multiple high-volume use cases, optimized for token efficiency, and supporting large-scale semantic and full-text search across knowledge bases, resulting in **\$60M** in cost savings. The system elastically scaled and handled multiple bullruns, with peak consumption of **100M** tokens per minute
- Leveraged LLMs to enhance customer support; achieved a **65%** automated resolutions, expanded coverage to **18** new countries. Optimized costs to only **\$2K** per month through semantic caching, enabling the solution to gracefully handle traffic spikes of up to **10x**.
- Developed cheap, customizable and compliance focussed guardrails, evaluation frameworks, and audit trail for critical business processes in Finance and Legal operations.
- Led various LLM-powered initiatives to enhance employee productivity, including coding assistants, on-call operations, Slack helpers, and various Legal and Finance business process automations.
- Leveraging LLM-as-a-Judge for automated and scalable evaluations and auto tuning prompts.
- Developed plugins for seamless integration of in-house LLM endpoints across popular agent frameworks (LangGraph, LangFlow, CrewAI, Legion AI, Bedrock Agents), enhancing developer workflow.
- **Technologies:** *LLMs: AWS Bedrock, GCP Vertex AI, Azure OpenAI, Anthropic, Huggingface, RAG, CAG, LangChain, LangGraph, DeepEval, VectorStore, MongoDB, AWS Aurora, Opensearch.*

Amazon

Staff Software Engineer   *Mar 2020 - Oct 2023*
AWS EMR

- Optimized AWS EMR Spark performance via improved query planning, S3 interactions, and JVM memory optimizations, achieving significant performance gains in TPC-DS.
- Enhanced performance via prefetch and dynamic partition pruning, optimal join strategies.
- **Technologies:** *AWS EMR, Apache Spark, Trino, S3, DynamoDb, JVM*

Amazon Business Automation

- Modernized Amazon's fraud and anomaly detection framework at a global scale by integrating real-time streaming to GNNs, reducing detection/mitigation times from weeks to one day.
- Architected a global scale serverless Data Lake for interactive analytics and machine learning training, resulting in a 3x improvement in operational metrics and a 10x reduction in costs.
- Built large-scale knowledge graphs with **50B** vertices to power GNN models. Efficiently caching subgraphs to reduce feature store costs by **\$500k annually**.
- Revamped the account management ecosystem by optimizing billing and metering pipelines, reducing turnaround time by **70%**. Enhanced rule engines to improve performance and cut costs by **65%**.
- **Technologies:** *AWS (Neptune, Textract, Comprehend, RDS Postgres, EMR, Athena, Lambda, SQS, SNS, DynamoDB, Redshift, OpenSearch, SageMaker), Apache Iceberg, DistDGL, Apache Spark*

Domino Data Lab
Staff Software Engineer   *Nov 2019 - Mar 2020*

- Optimized real-time feature store's operational costs by 45%.
- Designed consumer facing REST APIs for MLOps control plane.
- Designed an infrastructure-as-code template for high-volume data ingestion, enabling easy and repeatable deployments across tenant accounts.
- **Technologies:** *Scala, Play, Python, MongoDB, AWS (Sagemaker, EMR, S3, Kinesis, Lambda, EKS)*

Rivigo
Staff Software Engineer   *Jul 2018 - Nov 2019*

- Architected a scalable data ingestion platform, enabling a centralized Data Lake to streamlined analytics and ML workflows.
- Developed cost effective serverless feature store, that accelerated ML model training and inference by standardizing features across teams.
- Reduced AWS infrastructure costs by **40%** through optimization of service flows, transitioning from real-time to batch processing, and implementing efficient DB schemas and access patterns.
- Engineered a predictive maintenance system for truck fleet management, using sensor data feeds, enabling tracking and proactive vehicle maintenance to reduce downtime.
- Designed a flexible mapping framework supporting multiple vendors (Google Maps, Bing Maps, Map My India), optimizing route calculations and reducing operational costs by **30%**.
- Led the transformation of a monolithic Commerce service into specialized microservices (Order, Payments, Wallet, Incentives, Fuel, Fastag, Billing).
- Implemented real-time payout system with automated retries and intelligent bank selection, alongside a comprehensive wallet system integrated with bank accounts, fuel vouchers, and Fastag top-ups.
- Developed a easy to customise, DSL-powered incentive engine enabling business analysts to define and trigger real-time coupon distribution, enhancing user engagement and operational efficiency.
- **Technologies:** *Java, Scala, Python, Go, AWS (RDS, EMR, Glue, S3, ECS, Lambda, SNS, SQS), MongoDb, Postgres*

Adobe
Computer Scientist 2   *Jul 2013 - Jul 2018*

- Led development of ML-based document digitization pipeline for Adobe Sense-AI cloud offerings, successfully implementing feature extraction and HTML transformation frameworks that enhanced document accessibility on mobile devices, culminating in an **US patent (US20190163733A1)** for conversion quality evaluation.
- Engineered critical performance optimizations in AEM Forms, implementing selective HTML loading and enhanced XFA PDF rendering engines, while developing comprehensive form control systems for both frontend and backend data manipulation.
- Developed multiple PDF and XFA-Form authoring and management features for the Adobe Document Cloud.
- **Technologies:** *AEM, XFA PDF, Adobe Sense-AI, Scala, Java, Python, C++, Apache Spark, RNN*

PATENT

- Conversion quality evaluation for digitized forms., US10621279B2, filed under Adobe.

EDUCATION

Indian Institute of Technology, Guwahati   *2013*
Computer Science and Engineering   *CGPA:* **8.92**

- **Project:** Parallel GPU-Accelerated Visibility Graph Construction and Shortest Path Computation for Polygonal Obstacles
- **Technologies:** *C++, Boost, CGAL, GPGPU, CUDA, OpenCL .*