Sunil Sarolkar

📍 Pune, India | ✉ sunil.sarolkar@gmail.com | +91-9823578815 | 🔗 LinkedIn / GitHub / Website

Lead Data Engineer | Data Architect | 15+ Years Building Petabyte-Scale Financial Data Systems

Results-driven Engineering Lead with 15+ years of experience building high-performance data pipelines and defining Data Strategy for global financial institutions (Deutsche Bank, HSBC). Expert in the **Python data science stack**, **Spark**, and **distributed systems**, with a proven track record of processing massive datasets (1.5B+ records) to drive financial modeling and decision-making. Leading cross-functional teams with a PG Certification in Computational Data Science from IISc Bangalore. Skilled in bridging the gap between quantitative strategy and scalable engineering implementation.

Core Skills

Languages & Frameworks: Python, Java, PySpark, NumPy, Pandas, SQL, Financial Models, Data Strategy | **Numeric Storage:** HDF5, Iceberg, Parquet, Avro. | **Big Data & Data Pipelines:** Spark, Hadoop, Hive, Airflow, Kafka, BigQuery | **Cloud & Infrastructure:** GCP, AWS, Kubernetes, Linux, Terraform, CI/CD, Computer Networking (UDP, Multicasting) | **AI/ML:** TensorFlow, PyTorch, MLflow, LangChain, Hugging Face, RAG Pipelines | **Databases:** Oracle, PostgreSQL, MongoDB | **Programming:** Python, Java | Expert in Python and Java ecosystem.

Professional Experience

- **Assistant Vice President (Engineering Lead)** | *Deutsche Bank, Pune* | *Jul 2015 – Present*

Lead Data Engineer for enterprise cashflow analytics. **Engineered and performance-tuned** enterprise data applications using Python, Java, Spark, Dataproc and Cloud Data Platforms (GCP, AWS) to support financial operations, managing large-scale data processing. Architected scalable Data Lakehouse solutions (GCP Dataproc) for Cashflow Analytics, processing over **10 billion records monthly**. Championed **CI/CD** and **TDD/BDD** practices to ensure pipeline reliability. Defined and executed **Data Strategy** for cloud migration to Kubernetes. Managed end-to-end infrastructure using **Terraform (IaC)**, reducing operational overhead.

Deutsche Bank Cashflow Analytics Pipelines – Engineered and performance-tuned GCP Dataproc pipelines processing over 10 billion records monthly for critical cashflow analysis, resulting in a 40% reduction in BigQuery storage costs through optimization.

Advanced Analytics & Data Accessibility –
Partnered with cross-functional teams to deploy ML pipelines. Enhanced **Data Quality** and accessibility for data scientists by automating cleansing workflows.
Structform (Open Source, 2024) – Developed a Transformer-based structured data transformation engine achieving 99% validation accuracy; integrated automated SQL correction using LangChain + Hugging Face to improve data cleansing efficiency.
Generative AI Test Automation (2025) – Created a LangChain + Hugging Face pipeline for automated JUnit test generation, significantly reducing manual QA effort by 60% for core financial systems.

- **Cognizant Technology Solutions – Software Engineer (2013–2015):**

Built position management systems for global securities infrastructure, requiring an understanding of capital markets data.

- **HSBC GLT – Software Engineer (2011–2013):**

Built HSBC Smart Client Framework using Java, C#, enabling **low-latency** trading interactions and traders and stakeholders to interact with each other for matching trades.
Trading Platform Integration – Integrated proprietary trading platform with DTCC and MarkitWire using FIXML for clearing and regulatory compliance at a global investment bank (HSBC GLT).

Education & Certifications

- **PG Certification in Computational Data Science (CDS) – IISc Bangalore (2024):** Coursework included advanced topics in mathematical algorithms, probability analysis, and quantitative methods for data science, directly supporting financial modeling capabilities.
    - **AI-Driven Indian Sign Language Translation (IISc Capstone) | Live Demo**
    - Developed a real-time translation pipeline using OpenPose for skeletal keypoint extraction and Long Short-Term Memory (LSTM) networks for temporal gesture recognition.Deployed the inference engine as an interactive web application using **Streamlit** on **Hugging Face Spaces**, enabling real-time video translation.
- Google Certified Professional Machine Learning Engineer Certification – 2025
- Google Certified Associate Cloud Engineer – 2023
- Diploma in Advanced Computing (DAC) – ACTS, Pune (2007)
- B.E. Electronics – Pune University, VIT College (2006)