

Sumit Kumar

Data Engineer | NIT Patna '21

 +91-7549980508

 sumit749284@gmail.com

 linkedin.com/in/smaxiso

Skills

Languages: Python, Java, Scala

Data: Spark, PySpark, SQL, Kafka, Kafka Connect, Debezium CDC, Airflow

Cloud: AWS (ECS, MSK, DynamoDB, DocumentDB, S3, Bedrock), GCP (BigQuery, Dataproc, Dataflow)

GenAI: OpenAI GPT, AWS Bedrock, LangChain, Prompt Engineering

Frameworks: Spring Boot, FastAPI, React

Tools: Git, Docker, Jenkins, Maven

Experience

Gen Digital (formerly NortonLifeLock)

Data Engineer

Dec 2024 – Present

Chennai, India

- Engineered custom **Java-based Kafka Connect SMT library** for Debezium CDC processing from AWS DocumentDB, achieving **100% data integrity** and eliminating critical DebeziumException errors across real-time document pipelines.
- Implemented multi-field validation and smart filtering in **Spring Boot** transaction normalization services, **reducing invalid data processing by 40%** and supporting Gen's Unified Data Model initiative on **AWS ECS/MSK**.
- Designed and deployed **Python event-driven microservice** on ECS for member data operations, integrating Kafka for event consumption, DynamoDB for persistence, with comprehensive **GDPR compliance** and GUID sync strategy.
- Architected **Financial Insights Provider** — a Python/FastAPI microservice with **DynamoDB caching (3x speedup)**, 6 insight generators (Fee Monitoring, Duplicate Detection), and intelligent merge-on-PATCH logic for cross-scan deduplication.

FINN

- Led development of a **full-stack module** within the financial wellness platform leveraging **OpenAI GPT** for behavioral analysis, proprietary emotion scoring intelligence, and **Monte Carlo simulations** for retirement planning.

PayPal (via TCS)

Data Engineer

Jul 2021 – Nov 2024

Bangalore, India

Data Migration Framework (Mar 2023 – Nov 2024)

- Developed scalable **ETL framework** for PayPal's data migration using Python, AWS, and BigQuery — achieved **20% faster migration and 30% improved scalability**.
- Built automated dashboards with **Matplotlib** for stakeholder visibility and deployed orchestration using **Airflow** with auto-generated DAG scripts.

Lynx Entity Linkage Optimization (Jan 2024 – May 2024)

- Optimized **Lynx entity linkage framework** using Scala/Spark and Google's APSS algorithm — achieved **35% better accuracy and 40% faster approximate nearest neighbor search** via LSH enhancements.
- Conducted comprehensive testing with ML algorithms (RPDBSCAN, K-Means) for similarity scoring validation.

On-Demand Merchant Reporting (Aug 2021 – Jan 2023)

- Built **on-demand merchant reporting system** with Python, Oracle DB, and GCP Dataproc — achieved **15% improved accuracy and 25% faster report generation**.
- Created end-to-end pipeline integrating report requests, authentication, and automated orchestration via internal Airflow (DALM).

Education

National Institute of Technology Patna

Bachelor of Technology in Computer Science & Engineering

2017 – 2021

CGPA: 8.0/10