

Sumit Kumar

Data Engineer

+91-7549980508

sumit749284@gmail.com

linkedin.com/in/smaxiso

Education

National Institute of Technology Patna

Bachelor of Technology in Computer Science & Engineering

2017 – 2021

CGPA: 8.0/10

Experience

Gen Digital (formerly NortonLifeLock)

Data Engineer

Dec 2024 – Present

Chennai, India

Real-time Data Streaming Pipelines

- Engineered a custom Java-based Kafka Connect SMT library to process Debezium CDC events from AWS DocumentDB, eliminating critical DebeziumException errors and ensuring **100% data integrity** for real-time document processing.
- Implemented multi-field validation and smart filtering systems in Java/Spring Boot transaction normalization services, **reducing invalid data processing by 40%** and preventing empty transaction arrays from downstream propagation.
- Architected enhanced data streaming services on AWS ECS using Kafka (MSK) to support Gen's Unified Data Model (UDM) initiative, consolidating disparate data silos for critical smart alert generation.

Cloud-Native Services & Data Privacy

- Designed and deployed a Python-based event-driven microservice (on ECS) to manage member data operations, integrating Kafka for event consumption and DynamoDB for persistence with comprehensive GDPR compliance & GUID sync strategy.

Generative AI & Innovation (Hackathon)

- Led development of 'FINN,' a full-stack financial wellness platform during company-wide hackathon, featuring dual AI engines with **proprietary emotion scoring intelligence** and OpenAI GPT for behavioral spending analysis, plus Monte Carlo simulations for financial projections and retirement planning.
- Technologies used: Java (8, 17, Spring Boot), Python, Spring Boot, FastAPI, Kafka (MSK), Kafka Connect, Debezium, AWS (ECS, DocumentDB, DynamoDB), OpenAI GPT, React

Tata Consultancy Services - (Client: PayPal)

Data Engineer

July 2021 – November 2024

Bangalore, India

Data Migration Framework (Mar 2023 – Nov 2024)

- Developed a scalable ETL framework for PayPal's data migration using Python, AWS, and BigQuery, **reducing migration time by 20%** and improving scalability by **30%**.
- Built automated dashboards with Matplotlib for stakeholder visibility and deployed orchestration using Airflow with auto-generated DAG scripts.
- Technologies used: Python, AWS, GCS, BigQuery, Airflow, Matplotlib

Lynx Framework Optimization (Jan 2024 – May 2024)

- Optimized the Lynx entity linkage framework, achieving **35% improvement in data linkage accuracy** and **40% reduction** in approximate nearest neighbor search time through LSH algorithm enhancements.
- Leveraged Scala/Spark and Google's APSS algorithm for optimal performance in similarity scoring and conducted comprehensive testing with ML algorithms (RPDBSCAN, K-Means).
- Technologies used: PySpark, Scala, APSS, BigQuery, GCP (Dataproc, GCS), LSH

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

- Built on-demand merchant reporting system, **increasing data accuracy by 15%** and **reducing report generation time by 25%**.
- Created Python pipeline integrating report requests with authentication, Oracle DB validation, and GCP Dataproc processing, automated via DALM (internal Airflow).
- Technologies used: Python, SQL, Apache Spark, Oracle, GCP, Airflow, Dataproc

Forest Fire Detection System

- Developed a real-time **forest fire detection system** using Python-based ML algorithms and fuzzy logic, achieving **90% accuracy** in predicting fire likelihood and severity.
- Technologies used: **Python, Machine Learning, Fuzzy Logic**

Technical Skills

Programming Languages: Python, Java(8, 17, Spring Boot), Scala, C + +, Shell/Bash

Cloud Platforms: AWS (S3, MSK, ECS, EMR, Glue, Athena, DMS, KMS), GCP (GCS, BigQuery, Dataproc, Dataflow)

Data Streaming: Kafka, Kafka Connect, Debezium, Real-time CDC Processing

Frameworks: Apache Spark, PySpark, Spring Boot, FastAPI, Django, React

Databases: DocumentDB, DynamoDB, MySQL, BigQuery, Oracle

Developer Tools: Git, Airflow, TeamCity, Jenkins, Maven

Specializations: ETL/ELT Pipelines, Data Lake Design, Machine Learning, Generative AI, GDPR Compliance, Data Quality Engineering