# Sumit Kumar

Data Engineer

☎ +91-7549980508 ✉ sumit749284@gmail.com in linkedin.com/in/smaxiso

## Education

**National Institute of Technology Patna** **2017 – 2021**
*Bachelor of Technology in* ***Computer Science & Engineering*** *CGPA: 8.0/10*

## Experience

**Gen Digital (formerly NortonLifeLock)** **Dec 2024 – Present**
*Data Engineer* *Chennai, India*

**Real-time Transaction Normalization & Scalable Data Lake Design**

- Contributed to the **real-time transaction normalization pipeline** leveraging **Kafka (AWS MSK), ECS, and Java (Spring Boot)**, ensuring high-throughput data processing for fraud detection and financial alerts.
- Enhanced the **Normalization Service** to support **unified model payload changes**, improving data consistency and compatibility across multiple vendors.
- Assisted in the design and implementation of a **scalable Data Lake** using **S3 Hudi** to efficiently store and manage unified transaction payloads.
- Optimized the ETL pipeline framework for extracting data from **DocumentDB**, applying transformation logic to align with business requirements, and storing it in **Parquet format with Snappy compression** for efficient querying in **Athena**.
- Developed a **Java-based test automation service** with **Maven, TeamCity, and TestFLO**, streamlining automated test execution and Jira integration.
- Technologies used: **Python, Java, Scala, Kafka (AWS MSK), ECS, EMR, TeamCity, Artifactory, Airflow, AWS, DocumentDB, Athena, DynamoDB, S3 Hudi**

**Tata Consultancy Services - (Client: PayPal)** **July 2021 – November 2024**
*Data Engineer* *Bangalore, India*

**Data Migration Framework** (Mar 2023 – Nov 2024)

- Developed a scalable **ETL Framework for Data Migration** for PayPal using **Python, AWS, GCS**, and **BigQuery**.
- **Reduced data migration time by 20%**, improving scalability by **30%**.
- Created a dashboard in Python using **Matplotlib** for snapshot tables, providing data trend visibility to stakeholders. Automated the sending of dashboards via email daily, weekly, monthly, and half-yearly.
- Deployed the ETL framework and dashboard automation using **Airflow** with DAG scripts. Built an automated framework for configuration and DAG script generation.
- Technologies used: **Python, AWS, GCS, BigQuery, Airflow, Matplotlib**

**Lynx Framework Optimization** (Jan 2024 – May 2024)

- Implemented optimizations in the **Lynx Framework**, resulting in a **35% improvement in data linkage accuracy and efficiency**.
- Optimized the **Locality-Sensitive Hashing** algorithm, reducing approximate nearest neighbor search time by **40%**.
- Conducted thorough testing of the framework's performance and similarity scoring using ML algorithms such as **RPDBSCAN, LSH,** and **K-Means**.
- Leveraged **Scala** and **Spark** frameworks, utilizing **Google's APSS algorithm** to achieve the best performance and accurate similarity scores in entity linkage.
- Technologies used: **PySpark, Scala, APSS (All Pair Similarity Search), BigQuery, GCP (Dataproc, GCS), LSH**

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

- Built on-demand merchant reports, **increasing data accuracy by 15%**.
- **Decreased report generation time by 25%**.
- Created a pipeline in Python to integrate report generation requests with the report engine, integrated Keymaker authentication, Oracle DB validation, and triggered Dataproc for report generation.
- Automated the process using **DALM (an internal Airflow app)** to trigger every 30 minutes and one hour.
- Developed SQL queries for data validation and deployed them into the **Rule Execution Framework (REF)** for automated data validation.
- Technologies used: **Python, SQL, Apache Spark, Oracle, GCP, Airflow, Dataproc**

**NIT Patna (Internship)** **May 2020 – July 2020**
*Data Science Research Intern* *Patna, India*

**Forest Fire Detection System**

- Developed a real-time **forest fire detection system** using **Python-based machine learning algorithms** and **fuzzy logic**.
- Achieved an **accuracy rate of 90%** in predicting the likelihood and severity of forest fires.
- Technologies used: **Python, machine learning, fuzzy logic**

## Technical Skills

**Programming Languages:** Python, Java, C++, C, Shell/Bash

**Databases:** DocumentDB, DynamoDB, MySQL, BigQuery, Oracle

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

**Developer Tools:** Git, GitHub, CI/CD, Jenkins, Airflow, TeamCity, Artifactory, TestFLO

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

**Concepts:** Real-time Data Processing, ETL, Data Warehousing, Data Normalization, Data Lake Design, Transaction Processing, Machine Learning, Cloud Computing, Unix Systems, Generative AI, Agile Methodology, HDFS, Data Structures and Algorithms, Database Management, Operating Systems, Computer Networks