Sumit Kumar

Data Engineer

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

Education

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

Experience

Tata Consultancy Services **July 2021 – Present**
Data Engineer *Bangalore, India*

Data Migration Framework (Mar 2023 – Present)

- Developed a scalable **ETL Framework for Data Migration** for a leading global online payments company 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 **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, C++, C, Java, Shell/Bash

Databases: MySQL, BigQuery, Oracle

Frameworks: PySpark, Apache Spark, Django, React

Developer Tools: Git, GitHub, CI/CD, Jenkins, Airflow

Cloud Platforms: GCP (GCS, BigQuery, Dataproc, Dataflow, Data Catalog), AWS (S3, Lambda Functions, DMS)

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