Rohan Gupta

Noida, India | +91-8808219996 | rohangupta.00714@gmail.com | LinkedIn

SUMMARY

Experienced Data Engineer with 4 years of robust experience in Apache Spark, Kafka, ClickHouse, Superset, and cutting-edge Big Data technologies. Proven expertise in data pipelining, ETL, data visualization, job orchestration, batch and stream processing.

EDUCATION

B.Tech, Computer Science & Technology (2016 - 2020) Patna, India
8.21 National Insititute of Technology, Patna

WORK EXPERIENCE

InfoEdge India Limited — Noida, India

- Lead Engineer — Oct 2023 - Present

- **User Activity Central**
  - Created a user activity center comprising all major activities of daily active users on the platform for dynamic insights. Improved day 1 retention by 10% and day 8 retention by 12%.
  - Customizable filters using Jinja Templating to slice, dice, and analyze data across various cohorts which helped improve user experience in terms of engagement, impressions, retention, membership, etc.
- **Airflow Auto Healing Orchestration**
  - Enhanced system reliability by implementing auto-healing data-pipeline orchestration in Apache Airflow.
  - Applied external task sensors for monitoring task dependencies and ensuring seamless execution.
  - Reduced 90% manual intervention during failures by implementing DAG run idempotence, minimizing downtime and maintenance overhead.

- Senior Software Engineer — Jul 2021 - Sept 2023

- **User Funneling**
  - Capturing and storing clickstream data coming from different data sources such as apps and mobiles with the help of Java API.
  - Created funnels for Registration and Membership modules for identifying the dropouts and associated reasons with the help of Spark, Clickhouse, and Apache Superset.
  - Reduced registration drops by 15% in Android apps and 9% in IOS apps, and increased membership purchases by 7%.
- **Spam Identification**
  - Developed and deployed an ETL pipeline using Spark, Kafka, MongoDB Change Streams, and Maxwell for spam identification on Jeevansathi.com.
  - Upgraded from batch processing to real-time identification with Spark Streaming, blocking spammers within 5 minutes of profile creation.
  - Improved platform integrity by blocking 400-500 spam profiles daily, enhancing user trust and experience.

- Software Engineer — Jun 2020 - Jun 2021

- **Data Access Layer**
  - Built visualization for Jeevansathi to add BI over the data using open-source technologies like Superset, Clickhouse, S3, and CDC pipelines.
  - Enabling teams and PMs to run interactive queries, and create charts and dashboards for data visualization and analysis.
- **MySQL-MongoDB CDC Pipeline**
  - Designed and implemented CDC pipelines for MySQL and MongoDB ensuring accurate and timely data updates in the data warehouse using open-source technologies like Sqoop, Maxwell, Kafka, Spark, S3, and Clickhouse.
  - Gathering multiple data at a central location i.e. Clickhouse using the CDC pipelines has enabled informed decision-making due to interactive queries that drive growth alongside significant reduction in data storage.
- **Neighborhood Algorithm**
  - Implemented neighborhood algo for Jeevansathi.com using ETL pipelines and Apache Spark to eradicate the limitations of the existing system.
  - Increased recommendation by 2x, reduced computation time by 10x, and reduced high IO queries on the database.

SKILLS

Apache Spark, Clickhouse (Data Warehousing), Java, SQL, Airflow, Apache Superset, Apache Kafka, Maxwell's Daemon, Apache Sqoop, Python, Structured Streaming, Linux, DBMS, AWS S3, OOPs, OS, Git.

ACHIEVEMENTS

- Rockstar performer for Jan, Feb, Mar, and Apr in 2021 for Jeevansathi tech.
- Won Top Gun in Q1-Q2 in 2022 and Q1 in 2023 for best performance among all tech teams of Jeevansathi.