Subham Saha

Data Engineer | SQL, Spark, Python

saha.subham919@gmail.com +91 7001437293 Kolkata, India

PROFILE

Accomplished Data Engineer with 3.5+ years of experience in designing, developing, and optimizing scalable big data architectures. Proven expertise in ETL processes, data warehousing, and pipeline development using technologies like Teradata SQL, Spark, and AWS. Adept at leveraging programming skills in Python, Spark and SQL to transform complex data into actionable insights. Passionate about improving data accessibility and reliability to drive business intelligence and analytics.

EDUCATION

Bachelor of Technology, *Jalpaiguri Govt. Engineering College* 08/2016 – 10/2020
Electronics and Communications Engineering Jalpaiguri

Higher Secondary, *Laban Hrad Vidyapith* 05/2014 – 03/2016
Science Kolkata

Madhyamik, *Laban Hrad Vidyapith* 05/2008 – 03/2014
Kolkata

PROFESSIONAL EXPERIENCE

System Engineer, *Tata Consultancy Services* 02/2021 – 07/2024
Kolkata

- **Designed and implemented scalable data pipelines:** Developed and optimized ETL processes using Apache Spark and Teradata SQL to handle large volumes of structured and unstructured data.
- **Data warehousing:** Built and maintained data warehouses on platforms such as Data Lake, Teradata ensuring high performance and reliability for business intelligence needs.
- **Programming and scripting:** Utilized Python, SQL, and Shell Scripting for data processing, analysis, and automation tasks.
- **Cloud data infrastructure:** Migrated on-premises data systems to cloud platforms, leveraging cloud-native services for improved scalability and cost-efficiency.
- **Monitoring and troubleshooting:** Developed monitoring and alerting solutions for data pipelines and infrastructure, ensuring high availability and quick issue resolution.
- **Documentation and training:** Created comprehensive documentation for data workflows and provided training sessions for team members on new technologies and best practices.
- **Data transformation:** Applied complex transformation logic, including data cleaning, CDC and enrichment, using Python and SQL to ensure data quality and prepare it for analysis.
- **Loading and optimization:** Implemented efficient loading strategies into data warehouses (Teradata) and data lakes (AWS S3), optimizing for speed and reducing storage costs.
- **ETL automation:** Automated ETL workflows using Apache Airflow, scheduling and orchestrating data processing tasks to run reliably and reduce manual effort.

ADHOC PROJECTS

Airflow Migration 12/2023 – present

- Successfully led the migration of data processing workflows to Apache Airflow, streamlining task scheduling and orchestration
- Collaborated with cross-functional teams to analyze existing workflows and design scalable and maintainable workflows using Apache Airflow's DAGs (Directed Acyclic Graphs)

Data Hardening 05/2023 – 11/2023
Kolkata

- Implemented robust duplicate check sessions using advanced SQL queries, reducing data redundancy by 90% and enhancing overall data quality
- Spearheaded trend check sessions, leveraging SQL to identify and analyze data trends, providing valuable insights to inform strategic decision-making
- Successfully implemented re-startability sessions in SQL, enhancing the reliability of data processing workflows. This initiative resulted in a 70% reduction in data processing errors and minimized downtime

COURSES

Data Science and Machine Learning, *Bosscoder Academy* 05/2024 – present
kolkata, India

- Understand the fundamentals of data science, including data collection, cleaning, and preprocessing.
- Master statistical analysis and data visualization techniques.
- Gain proficiency in machine learning algorithms and their applications.

SKILLS

*** SQL**
Advance SQL, ETL, ELT, Spark SQL

*** Python**
Pandas,NumPy,Seaborn,Matplotlib

*** ETL tool**
Informatica.

*** Orchestration tools**
Uc4 and Apache Airflow.

*** Spark**
Spark Optimization, PySpark

*** Data Warehouse**
Teradata, Snowflake, Datalake.

*** Linux/Unix**
Shell Scripting for automating Tasks/ Jobs.

*** Cloud platform**
AWS [S3, Glue, EC2, Airflow, Redshift, Crawler, Lambda]
Azure [ADF, ADLS, Synapse, Storage, Data Bricks]

CERTIFICATES

- SQL (Advance) Certificate
- The complete SQL Bootcamp
- Bash Mastery: Become a Linux Power User

INTERESTS

Data Warehouses
Snowflake, Amazon Redshift

Data Science
AIML Models and Implementation

NoSQL databases
MongoDB, Cassandra