contact@shubhanshutrip.com
+91-9424592907

SHUBHANSHU TRIPATHI

Web: shubhanshutrip.com
LinkedIn: ShubhanshuTrip

TECHNICAL SKILLS

- **Programming Languages:** SQL, Python, PySpark
- **Azure Services:** Synapse Analytics, Databricks, Data Factory, SQL Database, Data Lake Storage, Logic Apps
- **Big Data Engineering:** Apache Spark, ETL Development
- **Databases:** Azure SQL Database, Microsoft SQL Server, MySQL, Oracle, PostgreSQL, SQLite

CERTIFICATIONS AND ACHIEVEMENTS

- Passed **"Microsoft Certified: Azure Data Engineer Associate" (DP-203)** certification exam (bit.ly/stdp203)
- Earned Gold badge (5 Stars) for SQL on HackerRank (bit.ly/sthrpf)
- Achieved 2nd runner-up position among 215 teams in i3i 2023, a Capgemini hackathon (bit.ly/sti3i)
- Passed "AWS Certified Cloud Practitioner" (CLF-C02) certification exam (bit.ly/stawsccp)
- Secured rank 6,888 among 1,08,495 candidates in GATE (CS) exam

PROFESSIONAL EXPERIENCE

Azure Data Engineer (Associate Consultant) **Capgemini** **March 2022 - Present**

End-to-end development of Enterprise Data Hub, Operational Data Store, Data Marts and Data Lakes

- Migrated on-prem big data ETL processes to cloud, by creating storage event and schedule triggered pipelines with medallion architecture in Azure using Python, SQL and PySpark
- Implemented change data capture (CDC) process to store transformed data with SCD Type 2 implementation
- Improved reusability by developing metadata-driven architecture to create dynamic pipelines, which selectively fetch data by joining required source tables and applying transformations, to generate PII-masked views and extracts as per business requirements
- Optimized pipelines by applying conditional activity executions to reduce average runtime by 38%
- Identified and automated the manual and repetitive tasks by developing dynamic Python scripts to generate SQL queries and create directory structures
- Implemented pre-load, data quality and data control checks
- Performed data cleaning and applied transformations on Parquet and CSV big data feeds
- Implemented status email notification functionality in pipelines using Azure Logic Apps and Web Activities
- Improved fault tolerance by identifying and covering multiple edge cases

Status and Metadata Reports Generation

- Dynamic real-time status and metadata tracking of ETL job extracts using Python
- Extracted metadata properties and row counts dynamically from DAT and TXT files

Miscellaneous

- Optimised Python programs to reduce average runtime by 23%
- Automated Excel macro runs by creating Python scripts, to email daily consolidated status reports

EDUCATION

Master of Technology (Computer Science and Engineering) **Birla Institute of Technology, Mesra** **CGPA: 8.06** **Duration: July 2018 to July 2020**

Thesis Work: Diabetes Prediction using Machine Learning (bit.ly/dbtspred)

- Achieved up to 81.6% accuracy in Diabetes prediction on Pima Indians Database with Random Forest classifier
- Applied and analysed the accuracies of "K-Nearest Neighbors, Support Vector Machine, Decision Tree and Random Forest" classification algorithms for diabetes prediction
- Achieved 7.04% improvement in the accuracy of Decision Tree classification algorithm for Diabetes prediction
- Predicted missing values present in the dataset using a set of "Linear Regression, Support Vector Regression, Decision Tree and Random Forest" regression algorithms
- Performed dataset balancing using SMOTE algorithm and then Feature scaling

PROJECT WORK

CoWIN Vaccine Notifier (bit.ly/covantf)

- Developed a real-time Covid vaccine availability tracker using Python to notify the user as soon as any desired vaccine is available on the CoWIN website, which helped more than 30 people to get Covid vaccines on time
- Implemented 4 dynamic filters on the vaccination calendar received as a JSON response from Co-WIN API