

## TECHNICAL SKILLS

---

- Python, SQL, PySpark (*Apache Spark*), Big Data, ETL (*Extract, Transform, Load*) development
- Azure (*Synapse Analytics, Databricks, Data Factory, Data Lake Storage, SQL Database, Logic Apps*)
- Amazon Web Services (*AWS*), Microsoft SQL Server, MySQL, Data Science, Machine Learning, Git / GitHub

## CERTIFICATIONS AND ACHIEVEMENTS

---

- Passed "Microsoft Certified: Azure Data Engineer Associate" (DP-203) certification exam ([bit.ly/stdp203](https://bit.ly/stdp203))
- Achieved 2<sup>nd</sup> runner up position amongst 215 teams in i3i 2023, a Capgemini hackathon ([bit.ly/sti3i](https://bit.ly/sti3i))
- Passed "AWS Certified Cloud Practitioner" (CLF-C02 / CLF-C01) certification exam ([bit.ly/stawsscp](https://bit.ly/stawsscp))

## EMPLOYMENT

---

### Azure Data Engineer

(Associate Consultant)

Capgemini

March 2022 - Present

#### Data Engineering

End-to-end development of Operational Data Store, Data Hub, 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 pre-load, data quality, and data control checks
- Performed data cleaning and applied transformations on parquet and CSV big data feeds
- Implemented change data capture (CDC) process to store transformed data with SCD Type 2 implementation
- Developed Data Marts by creating dynamic pipelines to selectively fetch data by joining multiple source tables and apply transformations, to generate PII-masked views and extracts as per business requirements
- Implemented dynamic pipeline status email notification functionality using Azure Logic Apps and Web Activity
- Optimized pipelines by applying conditional activity executions to reduce average runtime by 38%
- Identified and automated the manual tasks (like SQL queries creation) to save team's time and efforts

#### Software Engineering

Status and Metadata Tracking

- Dynamic real-time status and metadata tracking 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

July 2018 - July 2020

#### Thesis Work

Diabetes Prediction using Machine Learning ([bit.ly/dbtspred](https://bit.ly/dbtspred))

- Achieved up to 81.6% accuracy in Diabetes Prediction on Pima Indians Diabetes Database with Random Forest classifier
- Applied and analysed accuracies of "K-Nearest Neighbors, Support Vector Machine, Decision Tree and Random Forest" classification algorithms for diabetes prediction
- Achieved up to 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](https://bit.ly/covantf))

- Developed a Python notebook to notify the user, as soon as any desired Covid Vaccine is available on CoWIN website for booking
- Implemented 4 dynamic filters on the Vaccination calendar received as a JSON response from Co-WIN API
- Helped more than 30 people to get Covid Vaccines using this notifier