# Vishalbhai Barvaliya

Data Engineer
Toronto, ON | +1 (647) 894 2798 | vishalbarvaliya112@gmail.com | LinkedIn| GitHub | Leetcode

## WORK EXPERIENCE

**Data Engineer Associate** at Mundelbulb Technologies **:** Sept 2021 – Nov 2021

- Maintained data pipeline up-time of 99.06% while ingesting streaming and transactional data across different primary data sources using Spark, Databricks, Azure Data Factory, ADLS2, Blob Storage, Azure SQL Database, and Python.
- Automated and optimized ETL processes across millions of records, which reduced manual workload by 29% monthly.
- Ingested data from disparate data sources using a combination of SQL, Event Hub, and Azure Stream Analytics using Python to create data views to be used in PowerBI.
- Created monitoring alerts for data pipelines that improved the uptime of the network by 17% year over year.
- Created python library to parse and reformat data to reducing error rate and make code even more cleaner to read.
- Optimized PySpark Transformations using Adaptive Query Executions and partition pruning by enabling AutoBroadcastJoin to reduce network traffic and to increase effiency of our spark job in Azure Databricks.
- Wrote SparkSQL complex queries as per business logic to find business KPIs (average transaction by month, year in each city) and visualize it using PowerBI.

## SKILLS

- **Programming languages and Databases:** Python(Pandas, Numpy, Matplotlib), SQL, Scala, MySQL, PostgreSQL.
- **Big Data Frameworks:** Hadoop, Spark-Scala, Pyspark, Hive, Kafka, HDFS, Sqoop, Airflow.
- **Cloud Services(Azure):** Data Factory, Databricks, SQL Database, Synapse Analytics, Blob Storage, ADLS2.
- **Data Visualization Tools** : PowerBI, Matplotlib, Seaborn
- **Other Technical skills:** ETL/ELT, Data Warehousing, Data modeling, PowerBI, Data Structures and Algorithms.
- **Non-Technical Skills :** Strong verbal and written communication, Teamwork, Collaboration, Presentation.

## CERTIFICATIONS

- **Microsoft Certified**: Azure Data Engineer Associate
- **Databricks Certified** Associate Developer for Apache Spark
- **Databricks** Lakehouse Fundamentals
- **Astronomer Certification** for Apache Airflow Fundamentals

## PROJECTS

**Azure Data Engineering on F1 Racing Dataset** (Azure Cloud Services, Azure Databricks) Sept-2022

- Created an Automated system that can handle the hybrid load (full + Incremental) Using ADF, Databricks, PySpark, SparkSQL, and CLI which reduced manual data ingestion and processing by more than 70%.
- Ingested data from multiple sources into Azure Data Lake to make it ready for further transformations.
- Used PySpark to distribute data processing on large streaming datasets, for cleaning and transforming as per business logic and saved it in parquet to save memory.
- Developed dashboards to visualize data in different charts to get insights using PowerBI.
- Automated data ingestion pipeline to ingest data automatically periodically with incremental load.

**Azure Data Engineering on Covid-19 Dataset** (Azure Cloud Services, Azure Data Factory) May-2022

- Created Scheduled Pipeline in Azure Data Factory to Ingest data from HTTP and Azure BLOB Storage into Azure Data lake gen 2 and mapped that data using Data-Flow Activity.
- Ingested data from multiple data sources in a single data pipeline using Azure Data Factory's LookUp and Copy Activity.
- Copy data From Azure Data lake gen2 to Azure SQL Database using Azure Copy-data Activity.
- Scheduled data Pipelines Using the services of Azure Data factory and managed CI/CD workflow.
- Designed and implemented a real-time data pipeline to process semi-structured data by integrating millions of records from multiple data sources using PySpark in Azure Databricks and Azure Data Factory.

**Movie Recommendation Engine using collaborative filterning (**Pyspark, Seaborn, Pandas, SparkSQL) Dec 2022

- Developed ETL pipeline to extract 25 million records from movie lens dataset from various data formats into spark data frames using PySpark.
- Implemented search engine to our dataset to search for particular movie from thousands of records by movie name in just few milliseconds.
- Finally, created recommendation engine using scikit learn to implement collaborative filtering in project to provide best recommedation for movies to watch based on users favourite movies.
- Used partition pruning and adaptive query execution techniques to reduce shuffle and sort operations to increase efficiency and speed of data pipeline by almost 1/3 times.

## EDUCATION

- Lambton College, **Big Data Analytics** (Currently Enrolled) Jan 2022 – Aug 2023 | Mississauga, ON
- SSCCS College, **Bachelor of Computer Applications(BCA)** 2016 – 2019 | Gujarat, India