# Shashank Vankadari

(610) 570-6968 | shashank.vankadari@gmail.com | LinkedIn | GitHub

## SUMMARY

Data Engineer with 4 years of experience in developing and deploying data pipelines at scale for serving data warehouses and analytics platforms. Expertise in Python, Spark, SQL, ETL / Azure Data Lake, Power BI and Azure DevOps. Collaborated in a team of 7 and delivered cross-border projects in an agile framework, engaging in regular communication with clients during bi-weekly sprints.

## PROFESSIONAL EXPERIENCE

**Analytics Engineer | Lehigh University, USA** **June 2023 – August 2023**

- Engineered custom MICS survey of 8 datasets for over 50,000 individuals using Python and stored in MySQL for analysis.
- Confirmed the link between Water, Sanitation, and Hygiene insecurity and LSI through K-means clustering in Scikit-learn.
- Identified a significant correlation between water quality and LSI indicated by an adjusted R-Squared value of 0.86.

**Data Engineer | Infosys Limited, India** **March 2019 – July 2022**

**Healthcare client project**

- Orchestrated the development and ongoing maintenance of robust pipelines using Data Factory for ingesting data from 5-6 disparate sources, loading it into the raw data layer of ADLS Gen2 and transforming data using Databricks Spark jobs resulting in a 30% reduction in data ingestion time and enhanced data integrity.
- Leveraged Databricks Unity Catalog for managing data assets and performed dimensional modeling and achieved an average data loading speed of 1 GB per minute into Synapse warehouse.
- Supported the migration from on-premise warehouse to Snowflake, resulting in annual cost savings of $278,000 and a 14% increase in performance.
- Streamlined CI/CD processes with Ops team using Azure DevOps, achieving a 40% reduction in deployment time.
- Communicated with clients to gather future requirements to build 4+ pipelines for analyzing technical issues.

**Payment Hub project**

- Boosted data retrieval efficiency by 30% by automating data scraping with Python-based framework using Selenium web driver.
- Automated data transfers using Sqoop and Airflow, resulting in a 2x reduction in data transfer time between MySQL and Hive.
- Authored task definitions in Python for data transformation and loading into Hive and orchestrated them with Airflow jobs to meet Service Level Agreements (SLAs).
- Led performance tuning initiatives that optimized queries to yield a 5x improvement in data processing speed.
- Empowered parallel processing and concurrency by defining Airflow task groups which reduced latency by 20% with an impressive 97% execution success rate, improving workflow efficiency.
- Collaborated with DevOps to build a Docker image for production and integrated with Splunk for monitoring and alerts.

## TECHNICAL SKILLS

**Big Data Technologies:** Hadoop, Spark, Hive, Sqoop, HBase, Airflow, Kafka

**Cloud Stack:** Snowflake**,** Azure Cloud Platform - Data Factory, Synapse Analytics, ADLS Gen2, Databricks, Cosmos DB
**Languages:** SQL, Python, R, Shell Scripting
**Libraries:** Pandas, NumPy, Matplotlib, Plotly, Ggplot2, OpenCV, MLlib, Scikit-Learn, SciPy, Spark NLP, Streamlit
**Databases:** Microsoft SQL Server, SSIS, SSRS, MySQL, HBase, Cassandra
**Deployment & Project Management Tools:** GitHub, Azure DevOps, Docker, CICD, JIRA, Confluence
**Reporting Tool:** Power BI

## EDUCATION

**Master of Science in Data Science***,* Lehigh University, USA **December 2023**
**Courses:** Algorithms for Data Science, Optimization Techniques, Data Mining, Statistical and Machine Learning, Big Data Analytics.

**Bachelor of Technology in Electronics and Communication Engineering,** GITAM University, India **May 2018**

## CERTIFICATIONS

**Power BI Data Analyst Associate**, Microsoft **| May 2024 – April 2025**
**Azure Databricks,** Coursera **| March 2024**
**Career Essentials in Generative AI,** LinkedIn **| February 2024**