# SAMATHA YEDDULA

**Data Engineer**

346-434-8192 | yeddula.samatha@gmail.com
https://www.linkedin.com/in/samatha-y/
2126 Stella St Apt 3, Denton, TX - 76201

## Professional Summary

Highly skilled Data Engineer with 4+ years of experience building and managing scalable data pipelines on Azure cloud (Data Factory, Databricks, Synapse, ADLS Gen2). Expertise spans SQL, Python, and PySpark development, data modelling, data warehousing, and creating dashboards with tools like Power BI and Tableau. Proven ability to deliver actionable insights from large data sets. Passionate about leveraging data to drive business decisions.

## Skills

- Hadoop, HDFS, MapReduce, Hive, Apache Spark, Spark streaming, Snowflake, Kafka
- Azure Data Factory, Azure Databricks, Azure DevOps, Azure Stream analytics, Azure Synapse, ADLS Gen 2, SSIS, SSAS
- SQL, PLSQL, MySQL, Python, PySpark
- Power BI, Tableau, SSRS
- Git, GitHub, Eclipse, SSMS, Visual Studio, Microsoft Excel, MS Word, draw.io
- Excellent communication, Detail oriented, Problem solver, People Management

## Work Experience

**Azure Data Engineer: Pentair Inc** **Aug 2023 – Present**

- Designed, built, and automated batch and streaming data pipelines using Azure Data factory, Azure Databricks, Event hubs, Data Lake, and other azure services processing over 1 TB of data daily.
- Leveraged Azure Data factory activities like Metadata driven copy activity, foreach activity, copy activity, lookup activity, databricks activity, stored procedure activity, data flow activity etc., for creation of dynamic pipelines.
- Implemented a star schema architecture for building the data warehouse, resulting in a 30% improvement in query performance and Medallion Architecture (Bronze, Silver, and Gold layer) using Azure Databricks.
- Made use of SQL, Python (NumPy and Pandas) and PySpark scripts in databricks for cleaning and manipulation of data and used Delta Live Tables, Auto Loader in Databricks to progressively load streaming Cloud Files.
- Utilized Azure Repos for version control, Azure DevOps to deploy the data pipelines and to automate the CI/CD process, and Azure Boards for managing and tracking work items throughout the development cycle.
- Created scheduled and tumbling window triggers for time-based ingestion and transformation of data pipelines.
- Engaged in Agile scrum meetings, including daily stand-ups and globally coordinated Planning, to ensure effective project management and execution.

**Graduate Teaching Assistant: University of North Texas** **Aug 2022 – May 2023**

- Assisted professors in teaching data modeling and data visualization courses for master's students using tools such as draw.io, Lucid chart, Tableau and Power BI.
- Provided expert support in SQL querying, optimization, database operations, troubleshooting, dashboards, and data analytics ensuring students proficiency in database management.
- Successfully achieved 98% pass percentage of these sections by organizing quizzes, grading, giving apt feedback, and clearing doubts regarding SQL code and Tableau visualizations.

**Junior Data Engineer: Tata Consultancy services** **Apr 2019 – Sep 2021**

- Created data pipelines in Azure Data Factory (ADF) using Linked Services, Datasets, and Key vault to efficiently extract, transform, and load data from and to various sources/services such as API's, legacy systems, relational databases, Blob storage, Snowflake, and Azure SQL Server.
- Copied data with Change Data Capture from ERP (SAP/QAD) systems into ADLS as parquet format and executed merge, upsert, to implement SCD type 2 in transformation logic.
- Responsible for creating fact, dimension, staging tables and other database objects like views, stored procedures, functions, index, materialized views, and constraints.
- Created SSRS and Power BI dashboards and employed DAX functions.

- Developed ETL packages using SSIS for data extraction, data processing and data transformation, and loading and automated reporting and SSAS cube refreshes by scheduling the SQL Agent jobs.
- Achieved performance tuning by partitioning, indexing and wrote complex SQL queries including DDL, DML, triggers, CTEs, and window functions and implemented data quality checks and cleansing techniques for data accuracy.
- Proficient in handling errors and events, employing techniques such as precedence constraints, breakpoints, checkpoints, and logging.
- Used Git as a version control tool to maintain the code repository.
- Worked in a collaborative environment with data architects, data analysts, modelers, virtualization teams to gather requirements, design data workflows, and implement scalable data solutions.
- Knowledge of various phases of software development life cycle (SDLC) including requirements for gathering design, development, testing, deployment, and maintenance.

## Academic Projects

**Heart Disease Prediction**

- Developed a robust model for predicting heart disease risk, achieving an impressive accuracy of 90.32%, with key risk factors identified through correlation analysis.
- Evaluated multiple machine learning algorithms including Logistic Regression, KNN, Random Forest, and Decision Tree Classifier, noting that accuracy peaked at a 90:10 split ratio.
- Optimized Logistic Regression hyperparameters using Grid search to achieve peak accuracy.

**Visualization of Boston Housing Prices**

- Utilized Tableau to visualize the relationship between diverse housing features (e.g., number of bedrooms, floors, roof and heating type, kitchen design) and housing prices in Boston.
- Employed a variety of Tableau graphs such as line graphs, box and scatter plots, and heat maps to effectively illustrate trends, distributions, and correlations among different factors influencing Boston housing prices.

## Education

**University of North Texas**
Masters in Data Science — Aug 2021 – May 2023

**G. Pulla Reddy Engineering College**
Bachelors of Electronics and Communications Engineering — Jul 2015 – Apr 2019

## Certifications

- Certified as Azure Data Engineer Associate from Microsoft (DP 203) CERTIFICATE
- Databricks Certified Data Engineer Associate CERTIFICATE
- Certified in Azure Fundamentals from Microsoft (AZ 900) CERTIFICATE
- Databricks Accredited Lakehouse fundamentals CERTIFICATE
- Databricks Accredited Generative AI fundamentals CERTIFICATE

## Achievements

- Received On the Spot Award from TCS for my dedication and commitment.
- Received On the Spot(team) award from TCS for focus and outstanding contribution to the organization.