MITHUN DAMA

Jersey City, NJ | (551) 2859817 | mithundama.de@gmail.com | LinkedIn

Big Data Engineer & Educational Analyst with over **5** years of experience in developing large-scale, data-heavy applications, primarily in the **financial services industry**. Proficient in Python, SQL, and azure cloud services. Demonstrated expertise in data platform design, ETL processes, and data visualization. Seeking to leverage my skills and experience as part of the data strategy team.

TECHNICAL SKILL

Languages: Python, SQL, Py-Spark
Tools & Databases: Git, Power BI, Snowflake, Power Query, MS Excel, Spark, ETL/ELT, MSQL, SQL, NoSQL, API.
Machine Learning Theory: Supervised & Un-Supervised Learning.
Project Management: Agile – Scrum & Jira, Azure DevOps
Data Science Skills: Data Preprocessing, Data Transformation, Data Modelling, Data Analysis, Data Visualization, NLP
Cloud: Azure – SQL Database, Data Factory, Blob Storage, Azure Databricks, ADLS, VPC, Azure Synapse Analytics.

PROFESSIONAL EXPERIENCE

Data Engineer | IntelliLink Technologies, Princeton, New Jersey **Jun 2023 – Sep 2023**

- Extracted data from different databases and dumped it into Azure Data Lake Zen-2 using Azure Data Factory.
- Created three containers in Data Lake: Bronze (raw data), Silver (semi-transformed data), and Gold (transformed data).
- Connected Azure Data Bricks with Data Lake to transform data and used Synapse Analytics to create databases for transformed data.
- Generated reports using Power BI for transformed data, focusing on transactional data.
- Enhanced ETL processes, reducing data processing times by 20% and improving data quality.
- Integrated disparate data sources with a unified data warehouse in Snowflake, ensuring data consistency and reliability.
- Collaborated with data scientists, analysts, and business stakeholders to understand data requirements and documented data pipeline architectures and best practices.

Big Data Engineer | Yocket, India **Aug 2019 – Aug 2022**

- Designed and implemented data pipelines using Apache Spark to transform and process large-scale financial data.
- Extracted and transformed financial data from company databases using SQL, ensuring consistency and accuracy.
- Utilized Azure Data Factory and Azure Databricks for efficient data pipeline creation and management.
- Created and maintained tables in Snowflake for a centralized and scalable data warehouse, ensuring seamless data integration from various sources, including ADLS and Blob storage.
- Developed monthly financial forecasts using SQL and Excel, guiding strategic planning and optimizing budget allocations across departments.
- Created real-time dashboards in Power BI, presenting complex financial information in a user-friendly format, and connected Power BI to Azure Databricks to provide insights into key financial metrics.
- Applied statistical models to evaluate financial trends and assess risk, contributing to a 15% increase in resource allocation efficiency, directly impacting the organization's ability to support more students globally.
- Led a project to optimize the scholarship allocation process by analyzing historical financial data and student success rates, improving scholarship distribution efficiency by 20%.
- Collaborated with data scientists and analysts to understand data requirements and deliver actionable insights, documenting data processes, pipeline architectures, and best practices for future reference and scalability.
- Created various stored procedures to load the data from multiple sources to a single destination, ensuring data quality, reliability, and performance of big data systems.
- Contributed to the continuous improvement of the data infrastructure, ensuring data quality, reliability, and performance of big data systems.

Educational Analyst | Toppr, India **Jul 2018 – Aug 2019**

- Extracted data from various sources, transformed it using M language in Power BI, and efficiently loaded it into Power BI for reporting and analysis, developing and maintaining robust ETL pipelines.
- Utilized DAX measures in Power BI to create complex calculations and insightful visualizations, building comprehensive dashboards to present educational data.
- Created and managed databases using Azure services, ensuring data integrity and scalability, and set up and managed Azure Databricks clusters for data processing and analysis.
- Integrated Azure Databricks with Power BI for seamless data solutions, enabling real-time data updates and interactive visualizations.

- Implemented CI/CD pipelines to automate data workflows, ensuring efficient data pipeline management and timely updates.
- Collaborated with cross-functional teams to gather requirements and deliver tailored data solutions, documenting ETL processes, data models, and best practices.
- Led initiatives to improve data quality and streamline data operations, managing projects from conception to deployment and resulting in enhanced data accessibility and actionable insights for stakeholders.

EDUCATION

Pace University, NY
Master of Science in Data Science **May 2024**
Relevant Coursework: Data Mining, Machine Learning, Deep Learning, NLP, SQL, DBMS, Math & Stats, Artificial Intelligence (AI).

Kuppam Engineering College, Kuppam, India
Bachelor of Engineering **July 2018**

PROJECTS

Language Identification: -

- Worked on the NLP problem of Language Identification consisting dataset of 6,872,356 sentences and 328 unique languages to identify.
- Performed preprocessing on the large dataset of sentences and converted them to character-n grams to reduce the feature space and provide inputs to the Neural Networks as well as the models.
- Attempted to solve this NLP problem using a variety of methods including Naïve Bayes (giving accuracy-83.6%), Neural networks (accuracy-97.21%), and the BERT Multilingual base model(accuracy-99.2%).
- Conducted extensive hyperparameter tuning for the neural network models to optimize performance and reduce overfitting.
- Implemented model interpretability techniques to understand and visualize the contribution of different features in the language identification process.

Guest Profiling in Hotel Industry: -

- Processed and cleaned a large and complex dataset consisting of 1 million samples, including both real and generated dummy data, to track guest behavior.
- Employed Logistic Regression to create a classification model for predicting customer information and last review's attributes.
- Applied Multiple Linear Regression, Random Forest Regression, and Decision Tree Regression, comparing these models based on accuracy, AIC values, and significant variables.
- Built a Streamlit application for the predictor, utilizing machine learning and Python, and connected Azure Databricks for data transformation to expedite processing.