SETHUKUMAR . M

Data Engineer

USA | +1 857-313-5504 | sethukumarmogarala72@gmail.com | LinkedIn

Summary

Experienced Data Engineer with 4+ years specializing in architecting robust data platforms using **SQL**, **Python**, and **Big Data technologies**. Expertise in **AWS, GCP, CI/CD,** and advanced data visualization with **Tableau** and **Power BI**. Skilled in **Hadoop, Hive, Spark**, and machine learning frameworks like **Scikit-Learn** and **TensorFlow**. Led projects in automating ETL pipelines, enhancing efficiency by 60%, and ensuring data integrity through advanced security measures.

Experience

Data Engineer | Doublene, USA — Apr 2023 – Present

- Worked on developing **Pyspark** script to encrypting the raw data by using **Hashing algorithms** concepts on client specified columns, enhancing data security and confidentiality measures by 40%.
- Led the development of a scalable **big data processing** platform, handling daily data volumes of over 1TB, improving data processing efficiency by 50% and enabling real-time analytics capabilities.
- Create **Tableau** reports with complex calculations and worked on **ad-hoc reporting** using **PowerBI**, increasing reporting accuracy by 30% and enabling faster decision-making processes.
- Developed and maintained data pipelines using **Python**, **SQL**, and **PySpark** to perform **ETL** tasks from diverse sources like **AWS S3**, **Google BigQuery**, and **Azure Data Factory**, achieving a 60% reduction in data processing time.
- Performed **data analysis, data migration, data cleansing, transformation, integration, data import**, and **data export** through Python, optimizing data quality and achieving a 40% increase in data accuracy.
- Leveraged containerization and orchestration tools like **Docker** and **Kubernetes** to manage and scale data processing workflows effectively, accounting for 30% of job responsibilities.
- Developing Python-based APIs for revenue tracking and analysis to facilitate data-driven decision-making, improving revenue forecasting accuracy by 35%.
- Utilized **Scikit-Learn, TensorFlow,** and **Keras** to develop and implement machine learning models for predictive analytics, classification, and regression tasks, contributing to improved decision-making processes and achieving 90% accuracy in model predictions.

Data Engineer | TeCoventry, India — Jun 2018 – Dec 2021

- Engineered an end-to-end **ETL pipeline**, automating the process of data extraction, transformation, and loading into the reporting system, reducing manual tasks by 60% and error rates by 50%.
- Implemented **Waterfall methodology** for iterative development and rapid product delivery, ensuring project timelines were met with a 20% improvement in delivery speed.
- Supported the implementation of **Advanced Metering Infrastructure (AMI)** to enhance data accuracy and reliability in energy management systems, achieving a 25% increase in data accuracy metrics.
- Demonstrated a strong understanding of **AWS components**, utilizing **Redshift** to extract, load, and transform big data from various heterogeneous sources like **AWS S3, API**, and **Teradata**, achieving a 30% improvement in data processing efficiency.
- Carry out **Agile** and **SDLC methodology** to deliver end-to-end continuous integration/continuous delivery (**CI/CD**) pipelines, integrating tools like **Jenkins** and AWS for **VM provisioning**, resulting in a 40% reduction in deployment time.
- Deployed **Talend** for orchestrating data integration processes, including data quality checks, error handling, and monitoring, resulting in A 30% increase in data integration accuracy.
- Maintained and updated materials master data in **SAP ERP** system, ensuring data integrity and adherence to standards, achieving a 98% data accuracy rate.

Projects

Energy Consumption Prediction, Northeastern University Sep 2022 – Dec 2022

- Analyzed energy usage in different regions, processing 40 million data points by conducting feature engineering and optimizing hyperparameters through Bayesian Optimization, achieving a 25% improvement in model accuracy.
- Investigated various weather and consumption patterns at each location, leveraging Gradient Boost (XGBoost), Light Gradient Boost (LGBM), CAT Boost, and Linear Regression Models to achieve a 30% reduction in prediction errors.

Pothole Repair System Analysis, Northeastern University Jan 2023 – Apr 2023

- Designed a Pothole Repair System, combining MySQL and MongoDB databases to manage structured and unstructured data, reducing data retrieval time by 30%.
- Incorporated Python tools for streamlined data handling and reporting, enhancing pothole repair efficiency by 25%.

Regression Analysis on Demand of Bike Sharing and Rentals Jan 2024 - April 2024

- Conducted data cleaning, preprocessing, and EDA, achieving a 95% accurate dataset by reducing missing data
- by 60% and addressing 45% of outliers.
- Created Tableau dashboards for 50,000+ bike rental records, guiding resource allocation and marketing decisions. Identified ARMA as the optimal model with an RMSE score of 46.77.

NYPD Motor Collison Analysis Sep 2022 - Dec 2022

- Analyzed and pre-processed NYPD motor vehicle collision data of more than 100k rows, improving data accuracy by 20%.
- Visualized notable trends and patterns to provide interesting insights into data, increasing analytical depth by 30%.
- Complete a dashboard to visualize time series plots with varying boroughs, roadway users, and causality statistics, enhancing user interaction by 25%.

Skills

- **Programming Languages:** Python, SQL, PySpark
- **Big Data Technologies:** Apache Spark, Hadoop, AWS EMR, AWS Kinesis, AWS EC2, AWS S3
- **BI Tools:** Tableau, Power BI
- **Data Warehousing:** Amazon Redshift, Google Big Query, Azure Data Factory, Azure Databricks, Azure Synapse
- **Database Management:** MySQL, PostgreSQL, HBase, Cosmos DB
- **Data Skills:** Visualization, Data Modelling, Data normalization, Data Warehousing, Data Mining, Data Analysis
- **ETL Tools:** Apache Airflow, Talend
- **Tools and Software:** GitHub, MS Excel, MS PowerPoint, MS Word, SharePoint
- **Machine Learning:** Scikit-Learn, TensorFlow, Keras
- **Cloud Platforms:** AWS, AWS GLUE, AWS Redshift, AWS Lambda, Google Cloud Platform
- **Version Control:** Git
- **Data Modeling:** ERD, Dimensional Modeling
- **Operating Systems:** Windows, Mac
- **Soft Skills:** Critical Thinking, Write & Communication Skills, Presentation Skills, Problem-Solving, Team Workflows Leadership Skills, Project Management, USI Best Practice, Fast-Paced work skills, Integration skills, Risk Management, Marketing skills

Education

Master of Science in Data Analytics Engineering **Sep 2022 – May 2024**

Northeaster in University, Boston, MA

Bachelor of Science in Electronics and Communication Engineering: Concentration in Core Electronics **Jul 2018 – May 2022**

Vellore Institute of Technology, Amaravati, Andhra Pradesh, India