Raghu Vamsi

Data Engineer

Texas, USA | 469 296-8013 | raghu.vamsii98@gmail.com | **LinkedIn**

Education

Master of Science in Data Science – University Of North Texas, Denton, TX, USA

Bachelor of Technology in Civil Engineering - Jawaharlal Nehru Technological University, India

Skills

Methodologies:
SDLC, Agile, Waterfall

Programming Language:
Python, SQL, R

Packages:
NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, TensorFlow, Seaborn

Visualization Tools:
Tableau, Power BI, Advanced Excel (Pivot Tables, VLOOKUP)

IDEs:
Visual Studio Code, PyCharm, Jupyter Notebook

Cloud Platforms:
AWS (Amazon Web Services), GCP (Google Cloud Platform)

Database:
MySQL, PostgreSQL, MongoDB

Big Technology /Data Engineering Concept:
Apache Airflow, Apache Spark, Apache Hadoop, Apache Kafka, ETL/ELT, MapReduce, HDFS, Sqoop, Hive, NIFI

Other Technical Skills:
SSIS, SSRS, SSAS, Docker, Kubernetes, Jenkins, Terraform, Informatica, Talend, Amazon Redshift, Snowflake, Google Big Query, Data Quality and Governance, Machine Learning Algorithms, Airflow, Big Data, Advance Analytics, Statistical Methods, Data Mining, Data warehousing

Version Control Tools:
Git, GitHub

Operating Systems:
Windows, Linux

Summary

- Data Engineer with 4+ years of hands-on experience in data engineering, specializing in optimizing data processing, implementing ETL pipelines, and leveraging cutting-edge technologies for scalable solutions.
- Proficient in Apache Airflow, Apache Spark, Apache Hadoop, and Apache Kafka, utilizing these technologies to optimize data processing, enable real-time data streaming, and support large-scale analytics initiatives.
- Experienced in leveraging AWS (Amazon Web Services) for scalable storage, automated ETL job execution, and cost-effective data processing solutions, contributing to enhanced efficiency and reduced operational costs.
- Skilled in Python, SQL, and Scala for data processing, ETL operations, and exploratory data analysis, utilizing advanced packages and libraries such as NumPy, Pandas, Matplotlib, and Seaborn.

Professional Experience

Data Engineer | United HealthGroup, TX **Jul 2022 – Present**

- Implemented Apache Airflow to orchestrate and schedule complex data workflows, ensuring smooth execution of data pipelines and timely processing of healthcare data.
- Achieved a 30% improvement in data accuracy and a 25% increase in data throughput, ensuring more reliable analytics and reducing billing errors.
- Streamlined ETL pipelines for scalability, processing healthcare data with precision and speed, contributing to a 20% increase in claims processing efficiency.
- Implemented scheduled data ingestion and processing tasks using Apache Airflow, ensuring timely updates and availability of fresh data for analytics and reporting.
- Utilized AWS cloud services, including S3 for scalable storage and AWS Glue for automated ETL job execution, resulting in a 20% reduction in storage costs.
- Extracted data from multiple source systems S3, Redshift, RDS and Created multiple tables/databases in Glue Catalog by creating Glue Crawlers.
- Managed patient databases with PostgreSQL, optimizing queries and indexing for a 15% reduction in query response time.
- Optimized and tuned the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics.
- Employed Athena for interactive query analysis, Lambda for serverless computing, and Step Functions for orchestrating workflows, enhancing data processing capabilities and scalability.

Data Engineer | BirlaSoft, India **Jan 2019 – Jul 2021**

- Spearheaded a financial project at BirlaSoft using Apache Hadoop and Spark, achieving a 30% improvement in data retrieval speed with Amazon Redshift for data warehousing.
- Utilized PySpark for ETL processes, incorporating SQL scripts for data transformations, automating workflows, and reducing manual effort by 25% for streamlined and error-free financial data processing.
- Built real-time data ingestion pipelines using Apache Kafka on HDInsight for capturing and buffering user activity event streams before processing with Spark Structured Streaming.
- Implemented Power BI for interactive financial dashboards, connecting to SQL databases to facilitate SQL-based queries and contributing to a 25% improvement in decision-making speed for stakeholders.
- Performed incremental loads as well as full loads to transfer data from OLTP to the Data Warehouse of snowflake schema using different data flow and control flow tasks and provided maintenance for existing jobs.
- Conducted Exploratory Data Analysis using Python Matplotlib and Seaborn to identify underlying patterns and correlation between features.