Pragathi Gopishetty

Data Engineer

Maryland, USA | (732) 629-9387 | gopishettypragathi18@gmail.com | [LinkedIn] | [Github] | [Portfolio]

Summary

- Data Engineer with 4+ years of progressive experience in designing, developing, and maintaining robust data pipelines to integrate diverse structured and unstructured data sources.
- Demonstrated expertise in programming languages such as Python, SQL, and R, along with proficiency in essential packages including NumPy, Pandas, and SciPy.
- Experienced in leveraging cloud platforms such as AWS (EC2, S3, Lambda, Glue) and Microsoft Azure (Data factory, Synapse Analytics, Data Lake Storage, Azure SQL) for data processing, storage, and orchestration, with a focus on Azure Databricks for end-to-end data processing capabilities.
- Skilled in implementing data quality checks and validation rules to ensure data accuracy, completeness, and consistency, utilizing tools like Talend, DataStage, and QualityStage.
- Adept at developing resilient data models and schemas using technologies like Apache Hive, Apache Parquet, and Snowflake, facilitating efficient data storage, retrieval, and analysis.
- Proficient in visualization tools including Tableau, Power BI, and Advanced Excel, enabling insightful data visualization and analysis for informed decision-making.
- Experienced in version control using Git and GitHub, with a strong foundation in CI/CD pipelines and automation to streamline deployment processes.
- Committed to continuous learning and staying updated with emerging technologies and industry best practices in data engineering, ensuring the adoption of innovative approaches for solving complex data challenges.

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), Quick Sight
IDEs:	Visual Studio Code, PyCharm, Jupyter Notebook, IntelliJ
Database:	MySQL, PostgreSQL, MongoDB, SQL Server
Data Engineering Concept:	Apache Spark, Apache Hadoop, Apache Kafka, Apache Beam, ETL/ELT, PySQL, PySpark
Cloud Platforms:	AWS (EC2, S3, Lambda, Glue, Athena, SNS, RDS, EMR), Microsoft Azure (Data factory, Synapse Analytics, Data Lake Storage, Azure SQL), Databricks
Other Technical Skills:	Data Lake, SSIS, SSRS, SSAS, Docker, Kubernetes, Jenkins, Terraform, Informatica, Talend, Snowflake, Google Big Query, Data Quality and Governance, Machine Learning Algorithms, Natural Language Process, Big Data, Advance Analytics, Statistical Methods, Data Mining, Data Visualization, Data warehousing, Data transformation, Critical Thinking, Communication Skills, Presentation Skills, Problem-Solving
Version Control Tools:	Git, GitHub
Operating Systems:	Windows, Linux, Mac OS

Education

Master of Professional Studies in Data Science - University of Maryland Baltimore County, Maryland, USA
Bachelor of Technology, Computer Science and Engineering - MLR Institute of Technology, India

Certification

AWS Certified Solutions Architect Associate
Academy Accreditation - Databricks Lakehouse Fundamentals

Experience

Data Engineer | Kaiser Permanente, MD **Sept 2022 – Present**

- Developed and maintained robust data pipelines integrating diverse structured and unstructured data sources using tools like Apache Airflow and Apache Spark, resulting in a 20% reduction in data processing time.
- Extracted data from databases using Azure DMS/custom scripts while maintaining data integrity and performed end-to-end delivery of PySpark ETL pipelines on Azure Databricks to perform the transformation of data orchestrated via Azure Data Factory (ADF), achieving a 30% increase in data transformation efficiency.
- Implemented data quality checks and validation rules within Talend jobs to ensure the accuracy, completeness, and consistency of data.
- Integrated Power BI with Azure data services such as Azure SQL Database, Azure SQL Data Warehouse, Azure Blob Storage, and Azure Data Lake Storage (ADLS).
- Developed pipelines and data flows using Azure Data Factory (ADF), PySpark, and Data Flow within the Azure Databricks environment, emphasizing end-to-end data processing capabilities.
- Orchestrated Databricks Job workflows to seamlessly extract data from SQL Server and proficiently upload files to SFTP using PySpark and Python, resulting in a 40% decrease in data transfer time.
- Implemented CI/CD pipelines using Azure DevOps or Jenkins for automated deployment and version control of ETL (Extract, Transform, Load) processes.
- Leveraged JIRA's built-in features or third-party plugins to create chatops workflows, enabling team members to interact with Azure resources and services directly from JIRA channels or Azure Boards.
- Implementing high availability SQL Server solutions with SQL Server Failover Clustering, Database Snapshot, Replication, Log shipping, and Database Mirroring.
- Utilized DataStage and QualityStage for data profiling, cleansing, and deduplication tasks.
- Implemented data cleansing and transformation tasks within SSIS packages to ensure data quality and consistency during the migration process.
- Designed and deployed resilient data models and schemas utilizing technologies such as Apache Hive, Apache Parquet, or Snowflake, facilitating efficient data storage, retrieval, and analysis processes, resulting in a 30% improvement in data query performance.
- Optimized pipeline implementation and maintenance work using Databricks workspace configuration, cluster, and notebook optimization.

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

- Led the successful migration from Oracle to S3 using AWS DMS resulting in an annual cost savings of $650,000.
- Extracted insights from raw data using ETL processes with AWS Glue, Athena, Lambda, and Redshift, resulting in a 20% reduced user attrition.
- Leveraged Databricks and PySpark to architect robust data processing pipelines, optimizing queries with PySQL to improve efficiency. Enhanced CI/CD pipelines for Databricks, automating deployment processes seamlessly through git version control.
- Designed a scalable MySQL multi-tenant database architecture, allowing for efficient data management and analysis. Built RESTful APIs using Python Flask with Boto3 SDK to streamline data extraction from S3 buckets.
- Developed SSIS Packages optimizing ETL workflows with SQL server, reducing data processing time by 30% and enhancing loading phase accuracy by 25% and automated monitoring with Power BI, reducing manual efforts by 80%.
- Collaborated with stakeholders to define business objectives, analytical, problem-solving with attention to detail. Improved data integration, analysis, and test automation for customer data across 20+ clients.
- Analysed user experience by fine-tuning stored procedures and SQL queries for efficient data retrieval from approximately 4 million records accessed by 5000+ users worldwide.
- Orchestrated comprehensive infrastructure deployment utilizing Docker to build images, containerized with ECR and manifested using EKS. Showcased AWS infrastructure seamlessly using CloudFormation.