Sai Narayana Murthy Dontukurti

Data Engineer

+1 571-356-6474 | narayana.m@ajobguide.com | LinkedIn

Professional Summary

- Data Engineer with 3+ years of expertise in designing, building, and optimizing scalable data solutions, proficient in SQL, Python, AWS, data warehousing, data modeling, and ETL/ELT pipelines.
- Expert in leveraging AWS services (EC2, S3, RDS, Lambda, Glue, Athena, AWS Pipeline, Redshift) to build scalable and cost-effective data engineering solutions.
- Optimizing Data Build Tool (DBT) projects in a Snowflake environment by implementing incremental models, leveraging partitioning, tuning query performance, and reducing query runtime and cloud data warehousing costs.
- Experience in Database design, Data Modeling, Data Cleansing, and ETL Processes, with a deep understanding of both RDBMS (SQL Server, MySQL) and NoSQL (MongoDB, HBase) technologies, to design and implement solutions for diverse data needs.

Education

Master of Science in Data Science
George Washington University, Washington D.C

Skills

- **Programming Language:** Python, R, SQL, Spark SQL
- **Big Data Ecosystem:** Apache Spark, Apache Kafka, Apache Nessie, Hadoop, Hive, HDFS, MapReduce
- **Cloud:** AWS (EC2, S3, RDS, Lambda, Glue, Athena, AWS Pipeline, Redshift)
- **Visualizations:** Tableau, Power BI, Excel
- **Packages:** NumPy, Pandas, Matplotlib, Seaborn, PySpark
- **ETL and Tools:** SSIS, Informatica PowerCenter, Data Pipelines, Data build tool (DBT), Apache Airflow, Jenkins
- **Version Control & Database:** GitHub, Git, SQL Server, PostgreSQL, DynamoDB, MySQL, Snowflake

Experience

M&T Bank, VA | Data Engineer **Jan 2024 – Present**

- Established and maintained an ETL pipeline using Informatica PowerCenter to extract, transform, and load data from multiple sources into a data warehouse, ensuring data accuracy and consistency.
- Accomplished a complex data processing workflow consisting of multiple stages (data extraction, transformation, loading) using AWS Pipeline, improving data pipeline reliability and 40% reduction in data processing errors.
- Optimized Lambda functions for cost efficiency, achieving a 20% reduction in execution costs through code refactoring and best practices.
- Migrated data aggregation layer from legacy services to Snowflake unitizing Data Build Tool (DBT) models resulting in up to 70% cost savings and improved query performance.
- Orchestrated complex data pipelines with 3 stages using AWS Step Functions to automate data movement and transformation tasks, ensuring reliable data flow.
- Implemented a data warehouse on Databricks using Delta Lake, improving data query performance by 40% compared to the previous Teradata-based solution.

Fusion Software Technologies, India | Data Engineer - II **Aug 2021 – Jul 2022**

- Integrated Apache Airflow with AWS to monitor multi-stage ML workflows, with tasks running on Amazon SageMaker, and contributed to CI/CD solutions using Git and Jenkins for setting up and configuring the big data architecture on the AWS cloud.
- Created and managed 5+ ETL pipelines using AWS Glue, automating data extraction, transformation, and loading from various sources to a data warehouse.
- Built robust data quality checks and validation rules to ensure data integrity and accuracy throughout the Data pipeline process, reducing data errors by 50% and improving data reliability.
- Leveraged SQL and Alteryx to streamline data manipulation processes, resulting in a 20% improvement in data quality and a 30% reduction in processing times.
- Streamlined data ingestion and transformation processes by orchestrating multiple tasks and dependencies within Airflow DAGs.
- Enhanced AWS Redshift clusters for cost efficiency, achieving a 30% reduction in cluster costs, and maintaining query performance.

Cognizant, India | Data Engineer - I **Jan 2020 – Jul 2021**

- Executed ad-hoc queries and generated insights on petabyte-scale datasets using AWS Athena, providing real-time analytics to business stakeholders.
- Utilized Snowflake's query optimization features (materialized views, indexing) to improve query performance by 30% and reduce query execution time.
- Developed and optimized Apache Spark jobs for large-scale data processing tasks, including filtering, aggregation, and transformations using Spark SQL, DataFrames, and RDDs.
- Implemented checkpointing and state management strategies in Flink to ensure data integrity and recovery from system failures during real-time processing.
- Performed data sharing and collaboration using Nessie's branching and merging features, facilitating cross-functional data analysis.
- Employed ETL pipelines using Hive to extract, transform, and load data from various sources, increasing data ingestion rates by 25%.