RISHINDRA POPURI

DATA ENGINEER

+1 970-823-3396 | nagarishindra@careernb.com | LinkedIn

SUMMARY

- Data Engineer with 4+ years of experience across the data pipeline, from acquiring and validating large datasets (structured and unstructured) to building data models, developing reports, and utilizing visualization tools for impactful insights.
- Orchestrated complex data pipelines using DAGs (Directed Acyclic Graphs) in Airflow to automate data ingestion, transformation, and loading (ETL) processes.
- Excellent knowledge of AWS and Azure cloud services, including EC2, S3, RDS, Lambda, Glue, Athena, AWS Pipeline, Redshift, Azure DevOps, Azure Data Lake, Azure Data Factory, Azure Databricks with expertise in infrastructure management, storage, data warehousing, serverless computing, and automated deployment.
- Experience in building Spark applications (Python/PySpark) for large-scale data processing and improved processing speed compared to traditional methods.
- Ability to maintain the entire data pipeline infrastructure (Kafka, Snowflake, MongoDB) to ensure high availability and real-time data processing for fraud detection.

WORK EXPERIENCE

JPMorgan Chase & Co., TX | Data Engineer ***Nov 2022 – Current***

- Designed and implemented various Airflow DAGs (Directed Acyclic Graphs) to automate data pipelines, resulting in a 20% reduction in manual data processing tasks.
- Established real-time fraud detection pipelines using Spark Streaming and Scala to analyze high-volume transaction data, enabling immediate identification and prevention of fraudulent activity.
- Leveraged Databricks as a cloud-based platform for deploying and managing Apache Spark workloads, enabling efficient resource allocation and scalability for big data processing.
- Built data integration pipelines using Azure Data Factory to automate data movement and transformation between various data sources and sinks, reducing manual effort by 10%.
- Designed and implemented high-performance data warehouses on Snowflake for efficient storage, querying, and analysis of large and growing datasets, enabling scalability for future data needs.
- Applied data lake in Azure Data Lake Storage, improving data accessibility and collaboration for data scientists and analysts.
- Generated Spark applications in Azure Databricks that utilize in-memory processing for complex data transformations, achieving a 40% decrease in processing time compared to traditional batch processing techniques.
- Crafted and modified complex SQL queries, leveraging indexing and query optimization techniques to enhance database efficiency, leading to a 30% improvement in application response time.

Capgemini, India | Data Engineer ***Aug 2018 - Dec 2020***

- Increased data processing efficiency by 80% by migrating complex data transformations from AWS services to Azure Databricks within the data pipeline.
- Orchestrated data pipelines using Airflow to automate data ingestion, transformation, and loading tasks between various data sources (Kafka, HDFS) and data warehouses, ensuring reliable and scheduled data delivery for business intelligence dashboards.
- Enhanced Spark jobs to achieve 30% faster processing times by utilizing techniques like partitioning, data caching, and code optimization.
- Implemented materialized views and partitioning strategies in AWS Redshift to improve query performance by 15%.
- Built ETL pipelines in AWS Glue using PySpark to transform and load data into Redshift, reducing processing time by 25%.
- Executed serverless data pipelines using AWS Lambda to process real-time data streams, achieving a 10% reduction in processing latency compared to traditional batch processing methods.
- Created a Power BI dashboard to visualize key business KPIs, resulting in a weekly time savings of 1 hour on manual reporting.

SKILLS

- **Programming Language:** Scala, Python, R, SQL.
- **IDE's:** PyCharm, Jupyter Notebook.
- **Big Data Ecosystem:** Hadoop, MapReduce, Hive, Pig, HDFS, Spark, Kafka, PySpark, Apache Airflow, Zookeeper, Apache Flink.
- **Machine Learning:** Linear Regression, Logistic Regression, Decision Tree, SVM, K mean, Random Forest.
- **Cloud Technologies:** AWS (S3, EMR, EC2, Glue, Lambda, SDK, DynamoDB, Elasticsearch, QuickSight, Kinesis, Athena, VPC, Redshift), Docker, Azure (Data Lake, Data Factory, Databricks, Logic Apps, HDInsight, Synapse Analytics, Stream Analytics)
- **Packages & Data Processing:** NumPy, Matplotlib, Seaborn, TensorFlow, Plotly, PySpark, Data Pipelines, Jenkins
- **Version Control & Database:** GitHub, Git, SQL Server, PostgreSQL, MongoDB, MySQL, Snowflake
- **Operating Systems:** Windows, MacOS, Linux (Debian)

EDUCATION

Master of Science in Computer Science ***Aug 2022***

University of Texas at Arlington, Arlington, TX

CERTIFICATIONS

- Microsoft Certified: Azure Data Engineer Associate