JD: Data Engineer

About Us

Renan is a leading AI-model training and GenAI Applications development company. It partners with enterprises to increase efficiency and set them up for success through leading AI/ML-led solutions. Renan specializes in AI-led Analytics, Data Labelling and Annotation for training AI Models across 50+ languages from a skilled workforce across 30+ countries worldwide. Our founders have a total of 40+ years of experience in the industry and have collectively led large teams of 25,000+ employees across the globe in prominent organisations and startups like Amazon, HSBC, Trell (India's first social e-commerce platform), Coindelta (India's first cryptocurrency exchange) and Jar.

Our Products and Services include:

- **Friday** – AI Business Application for Data Analytics
- **Databrewery** – AI Model Training Platform
- **AI Consulting & Project Execution Practice**

Role Overview

As a **Data Engineer**, you will be responsible for designing, developing, and maintaining scalable data pipelines and infrastructure. You will work with structured and unstructured data, ensuring efficient data ingestion, transformation, and storage to support analytics, machine learning, and business intelligence needs. You will collaborate closely with data scientists, analysts, and software engineers to build robust and high-performance data solutions.

Key Responsibilities

- Design, develop, and maintain **ETL/ELT pipelines** to process large volumes of data efficiently.
- Work with **SQL and NoSQL databases** to ensure optimized data storage and retrieval.
- Develop and maintain **data lakes and data warehouses** using cloud-based solutions (AWS, GCP, Azure).
- Ensure **data quality, integrity, and governance** by implementing best practices for data validation and monitoring.
- Optimize **data workflows and performance tuning** to enhance query speed and system efficiency.
- Collaborate with **cross-functional teams** to integrate data solutions into various applications and services.
- Implement **real-time and batch processing** using tools like Apache Spark, Kafka, or Flink.

- Work with **cloud-based data services** (BigQuery, Redshift, Snowflake) for scalable and cost-effective solutions.
- Automate data pipeline deployment using **CI/CD and infrastructure-as-code** tools.
- Monitor and troubleshoot **data pipeline issues**, ensuring minimal downtime.

Required Skills & Qualifications

- **3+ years** of experience in data engineering, data architecture, or a related field.
- Strong proficiency in **Python, SQL**, and scripting for data processing.
- Experience with **big data processing frameworks** such as Apache Spark, Hadoop, or Flink.
- Hands-on experience with **ETL tools** like Apache Airflow, DBT, or Talend.
- Knowledge of **cloud platforms** (AWS, GCP, or Azure) and their data services (Redshift, BigQuery, Snowflake, etc.).
- Familiarity with **data modeling techniques**, database indexing, and query optimization.
- Understanding of **real-time data streaming** using Kafka, Kinesis, or Pub/Sub.
- Experience with **Docker and Kubernetes** for deploying data pipelines.
- Strong problem-solving and analytical skills, with a focus on performance optimization.
- Knowledge of **data security, governance, and compliance best practices**.

Preferred Qualifications

- Experience with **machine learning pipelines** and integrating data engineering with AI/ML workflows.
- Knowledge of **IaC tools** like Terraform, CloudFormation for automating infrastructure.
- Familiarity with **graph databases and time-series databases**.
- Previous experience in a **fast-paced startup environment** or working with large-scale distributed systems.