Data Engineer – Databricks Specialist

JD-0054

Key Responsibilities

• Design, develop, and maintain scalable data pipelines on Databricks using Spark, Delta Lake, Workflows, and DLT pipelines.

• Implement and manage Medallion Architecture (Bronze/Silver/Gold layers) with a strong focus on data quality, performance, and reliability.

• Build and optimize Structured Streaming pipelines for real-time and near real-time data ingestion and processing.

• Configure and tune Databricks clusters, jobs, and workloads to balance performance, scalability, and cost efficiency.

• Implement data ingestion from multiple sources (batch, CDC, streaming) across cloud data platforms, primarily AWS Databricks.

• Enforce data governance, security, and access control using Unity Catalog, schema management, and robust data-quality checks.

• Enable secure data sharing and collaboration through Delta Sharing and well-defined data contracts.

• Partner with analytics, BI, and ML teams to provision curated datasets and reusable data assets that support reporting, advanced analytics, and AI/ML initiatives.

Required Skills

• Strong hands-on experience with **Databricks** on cloud platforms, specifically **AWS Databricks**.

• Deep expertise in **Apache Spark** and **PySpark** for large-scale data processing.

• Proficiency in **Python** for data engineering, orchestration, and automation.

• Advanced **SQL** skills and solid understanding of **data modeling** (dimensional, relational, and analytical models).

• Practical experience implementing **Medallion Architecture** (Bronze/Silver/Gold) on Databricks.

• Hands-on experience with **Delta Lake**, **DLT (Delta Live Tables)**, and **DLT pipelines**.

• Strong knowledge of **Structured Streaming** for building real-time and incremental data pipelines.

• Experience with **Unity Catalog** for data governance, security, and metadata management.

• Expertise in **Delta Sharing** for secure and governed data-sharing capabilities.

• Experience with at least one major cloud platform (AWS required; GCP/Azure exposure desirable).

• Strong understanding of data ingestion patterns (batch, CDC, streaming) and performance tuning in Spark.

• Excellent analytical, problem-solving, and troubleshooting skills, with a platform-oriented mindset (reusable frameworks over one-off pipelines).

Good to Have Skills

• Knowledge of **data warehousing** concepts and tools (e.g., Redshift, Snowflake, BigQuery, Synapse).

• Familiarity with **orchestration tools** (e.g., Airflow, AWS Step Functions, Azure Data Factory).

• Understanding of **CI/CD** and DevOps practices for data engineering (e.g., Git, Jenkins, Azure DevOps, GitHub Actions).

• Exposure to **ML/AI workflows** and integration of Databricks with ML frameworks.

• Knowledge of **data cataloging** and **metadata management** tools beyond Unity Catalog.

• Experience with **monitoring and observability** for data pipelines (logging, alerting, data quality dashboards).

• Domain experience in **logistics, supply chain, or transportation** (aligned with ’ business domain).

Experience Required

10 to 15 years of overall experience in data engineering, with substantial recent experience focused on Databricks and Spark-based solutions.