Data Architect / Engineer Scope of Work: 2024 Enterprise Housing & Education

INNOVATE MEMPHIS
DELIVERING CIVIC SOLUTIONS

Overview

Innovate Memphis is seeking a skilled Data Architect/Engineer to lead the creation of our modern data infrastructure—comprising a Databricks Delta Lake, Data Catalog, and Monitoring Framework.

By architecting a scalable solution and engineering reliable data pipelines, you will impact thousands of students because this work is pivotal for understanding how Memphis's housing conditions influence student outcomes in education; It will enable partners to design data-driven strategies that address student chronic absenteeism and other critical community-specific challenges. Moreover, your work will be instrumental for future projects, including an "Eviction Early Warning System" that will allow Memphis Shelby County Schools to proactively provide support to students and families navigating housing instability.

In this role, you are not just building systems; you're uplifting impactful, community-empowering initiatives.

Contract Duration

This contract spans 8-12 weeks of dedicated development, plus two months of subsequent post-deployment support. We value work-life balance, offering schedule flexibility and potential contract extensions based on mutual agreement.

Location

We prioritize local talent but welcome remote candidates from anywhere within these United States.

Budget

Totaling $22,050, this contract allocates $18,900 for dedicated development work (180 hours at $105/hour) and $3,150 for post-deployment support. Full payment is contingent upon the achievement of KPIs and successful user acceptance testing, which will be agreed upon contractually.

Tools & Technology

- Data Sets: Student Information, Student Historical Addresses, Student Attendance, Student Academic Performance, Shelby County Addresses, Home Sales, Housing Eviction Filings, MLGW Utility Cut-Off Events, USPS Housing Vacancies, Assessor Land Bank, Assessor Land Use, Building Code Enforcement Records, County Building Permits, Environmental Court Cases (NPA), Comprehensive Housing Affordable Strategy values, U.S. Census at the Block, Zip Code, Tract, and County Level
- Storage: Microsoft Azure Blob, Databricks Delta Lake
- Data Integration: Databricks with custom workflows and notebooks in R, Python, or SQL
- Version Control: Git, GitHub

Stakeholders

Your close collaborators will include the Director of Data Midsouth, data analysts, researchers, and community partners like Seeding Success, Enterprise Community Partners, and Memphis Shelby County Schools.

Reporting

Bi-weekly virtual meetings with the Director and regular KPI assessments will keep the project aligned, proactive, and goal-centric.

Data Architect / Engineer Scope of Work:
2024 Enterprise Housing & Education

INNOVATE MEMPHIS

DELIVERING CIVIC SOLUTIONS

Goals & Objectives

Your primary goals include the effective implementation of a Databricks Delta Lake, robust data pipeline construction and orchestration within Databricks, comprehensive Data Catalog creation, developing complex queries, and ensuring strict data integrity, compliance, and governance. Additionally, you'll establish a Monitoring Framework for system reliability.

1. Databricks Delta Lake Implementation: Test and production environments include housing, geographic, and education data from 16 distinct structured datasets, each with a size less than 1GB.
2. Data Pipelines: Migrate existing housing data, followed by incremental batch loading of several data sets on a quarterly basis, utilizing API calls or Secure File Transfer Protocol (SFTP) methods from PostgreSQL databases.
3. Data Catalog Creation: Design to facilitate easy data discovery, access control, and efficient management through detailed metadata.
4. Complex Queries: Develop virtual views for Vacancy Index methodology and data sets for analyses.
5. Data Integrity: Ensure rigorous data integrity despite moderate source discrepancies through strict validation rules, error-checking mechanisms, and redundancy checks.
6. Compliance: Uphold stringent security standards, particularly adhering to FERPA student confidentiality laws.
7. Monitoring Framework: Provide complete system insights to track data quality, pipeline performance, and system health.

Key Performance Indicators (KPIs)

1. Data Completeness: All requisite data is present post-ETL.
   Report any instances of incomplete data, identifying affected tables, columns, and records.
2. Data Accuracy: Rate of records with discrepancies is less than 0.5% when compared to source data.
   Detail discrepancies discovered, potential root cause, and corrective action taken.
3. Low Pipeline Latency: Latency is less than 20 minutes from source to Databricks environment.
   Report average latency, instances exceeding KPI, and resolution details.
4. Query Performance: Average execution time is between 10 - 75 seconds.
   Share query profile to report progress and specify actions taken or planned.
5. Error Documentation: 100% of ETL errors should be documented with root cause and resolution steps.
6. Catalog Accuracy: Catalog entries reflect real-time changes and additions with 95% accuracy.
   Report automated validation accuracy of metadata and source data assets.
7. Monitoring Efficiency: Ensure the Monitoring Framework detects and reports issues proactively, maintaining system reliability and data integrity.

Data Architect / Engineer Scope of Work:
2024 Enterprise Housing & Education

Tasks & Deliverables

Planning & Design (2-3 weeks)

- Design the scalable Delta Lake infrastructure within Databricks.
- Initiate data ingestion/ETL workflows, Data Catalog structure, and Github for version control.
- Establish metadata tagging and management protocols within the Data Catalog.
- Implement the Monitoring Framework, leveraging Databricks and Azure functionalities (or third-party software like Data Dog).
- Collaborate on user requirements, and document system architecture and processes comprehensively (i.e., diagrams, in-line code, semantic code as documentation, where possible).

Development (4-6 weeks)

- Develop the Delta Lake and data ingestion pipelines, ensuring compatibility with diverse data sets.
- Create efficient, reusable scripts and conduct extensive ETL and security testing within Databricks notebooks.
- Engage with stakeholders for validation, data model testing, and exploring improvement opportunities.
- Generate optimized queries as views to support future statistical analysis.

Deployment (2-3 weeks)

- Conduct extensive testing, validating each component's functionality and performance.
- Train relevant stakeholders and users on how to use the Data Catalog, Delta Lake, and Monitoring Framework effectively.

Post-Deployment (2 months, flexible based on need)

- Offer dedicated post-launch support, addressing technical concerns promptly and refining system aspects based on user feedback.
- Estimated 15 hours or less per month.

How to Apply

Submit the following materials by November 17, 2023, 6:00 PM PST with the subject line "*Data Architect/Engineer Application*" to Kira Morin (kmorin@innovatememphis.com)

1. Your resume (pdf)
2. An outline of your strategy for the project tasks, including any innovative approaches
3. How your background prepares you for this role

Portfolios are optional but beneficial

About Innovate Memphis

Innovate Memphis is a 501(c)(3) nonprofit that develops and delivers bold solutions to some of the most pressing urban challenges in Memphis. Our organization advances programs that enhance the lives of Memphians and elevate the public good. We focus on a small number of challenges, gain deep insight and then apply innovation, demonstration, and incubation principles on a scale that reduces risk and finds a clear path to success. This model has proven results in youth gun violence, neighborhood economic vitality, and city service delivery.