New York University
Computer Science Department
Courant Institute of Mathematical Sciences

Database Systems Project Part IV
End-to-End Solution Integration and Data-Driven / Database Programming

Due Date: 12/22/22 – 11:59 pm ET

Course Title: Database Systems
Course Number: CSCI-GA.2433-001
Instructor: Jean-Claude Franchitti
Session: 11

1. Ongoing Project Background

Most enterprises today still rely on structured data stored in traditional relational databases and data warehouses. In addition to using these data sources, enterprises need to derive real-time insights to ensure business growth, "go digital", and drive business decisions that improve user experience and organizational excellence. To support the same, enterprises are designing and deploying additional data sources that manage large amounts of unstructured data to enable semi real-time big data analytics and create machine/deep learning and/or AI digital solutions.

The project initially focused on data stored in traditional relational databases that is typically used by insurance companies to conduct reporting and traditional business analytics. It is typically the case that many relational databases replicate metadata and related data in many parts of the enterprise. This drives the goal to establish an Enterprise Data Architecture (EDA) and to promote subsequent activities related to the integration of existing and new projects with the EDA. There are typically three separate efforts that are part of the creation of an EDA:

- Modeling – Creation of a diagram and/or blueprint that support the design of enterprise storage systems
- Operational Data Store (ODS) – Creation of physical database(s) that conform to the model
- Roadmap – Identify means to move applications / operations to integrate with the ODS

The first part of the project analyzed an existing logical model and led to the creation of a documented entity-relationship diagram using a mainstream software tool. The resulting model was partially validated against a set of business requirements and rules that were amended as/if needed. The second part of the project focused on the collection of unstructured data to drive business decisions meant to improve user experience and operational excellence. This resulted in the creation/generation and optimization of a logical database schema for the conceptual model created in the first part of the project. The resulting schema was then able to inter-relate structured and unstructured data, which resulted in an hybrid logical data model/data lake that was able to capture insights and drive decisions. The third part of the project focused on the creation and deployment of an optimized physical database model for the relational database

schema created in the second part of the project. It also focused on the creation of a machine learning model to perform analytics on the unstructured data collected in the second part of the project.

This final part of the project focuses on the implementation of an end-to-end program that updates the OLTP/ODS relational database created and optimized throughout the project to consider insights obtained by applying a machine learning model on unstructured data to help maximize operational excellence and business competitiveness. The end-to-end solution should support ongoing updates to unstructured data, as it is being mined, that may result in the need to re-train the machine learning model developed during the second part of the project. The overall solution should operate seamlessly and integrate the management of the hybrid data pipeline with minimal human intervention.

To meet the end goal of being able to drive business decisions that improve user experience and organizational excellence, it is necessary to keep refining the "end-to-end" reference architecture that was developed and improved upon in the second and third parts of the project. The reference architecture should span across the business, application, pyramid of knowledge (DIKW[1]), and infrastructure domains. As explained earlier, a reference architecture consists of foundational principles, an organizing framework, a comprehensive and consistent method to plan, deliver and operate business solutions, and an overarching governance. Governance is the set of processes and organizational structures that ensure conformity to the reference architecture principles, policies, and guidelines. With respect to the DIKW domain, metadata management, data quality, and data governance/intelligence are key ingredients that most enterprises need today to conduct business. Without these ingredients, as terabytes of structured and unstructured data flow into data lakes, it becomes extremely difficult to sort through their content and keep them from becoming unusable "data swamps". Enterprises also have to put in place robust management practices to secure their cloud and prevent data loss and leakage. DIKW governance today has many facets and includes such aspects as governing the lifecycle of data (e.g., data modeling). It must also safeguard enterprises against potential bias subsumed in socio-technical systems and limit the decision power of such systems to ensure fairness, accountability, and transparency.

2. End-to-End Solution Design and Implementation

The following steps should be followed to develop the end-to-end solution described in the previous section:

1. Select, design, and implement appropriate business use cases to support the creation of a data-driven workflow-based database application of your choice. For example, your business use cases may enable customers to obtain insurance quotes and policies once they select insurance products. In that case, insurance products and/or policy rates could become available or be updated on an ongoing basis.

2. Document your business use cases and the processes used by your application using a modeling notation of your choice. You should also document the design of your data-driven workflow-based database application.

3. Create a data-driven program module, using programming techniques and language(s) of your choice, that leverages the machine learning model that was developed and trained as part of the previous part of the project. You data-driven module should follow the design provided

[1] https://towardsdatascience.com/rootstrap-dikw-model-32cef9ae6dfb

in response to question 2 above. It should also manage the data pipeline for your unstructured data and enable seamless re-training of your machine-learning model upon changes to the corresponding source of unstructured data.

4. Using a database connectivity framework and programming techniques/language(s) of your choice, implement an end-to-end workflow-based application based on the design provided in response to question 2 above. Your application should integrate the use of the data-driven program module implemented in response to question 3 above.

   Note: extra credit will be granted for the use of an Object Relational Mapping (ORM) framework as part of the solution developed in response to this question.

5. Document your application and explain how your overall solution meets the various requirements set forth in this project part 4 specification. You should focus on making sure that the solution is "end-to-end" and integrates the gathering of data-driven insight computations and their integration with the OLTP/ODS database application that is made available to end-users. You should also explain how you optimized the various database queries; techniques to be considered for optimization purpose may include query optimization and/or denormalization as applicable. Finally, you should explain how optimizations were achieved as part of the overall solution implementation if you made use of an ORM framework as part of your solution.

6. Finalize your "end-to-end" reference architecture (RA) documentation. Please note that the reference architecture should span across the business, application, pyramid of knowledge (DIKW[2]), and infrastructure domains. In particular, you should clearly state the RA foundational principles and describe its organizing framework, the comprehensive and consistent methods applied to plan, deliver and operate business solutions, and the overarching governance. Please address the various aspects of data governance (set forth in Section 1 of this specification), which you believe are relevant to the enterprise goals and solution you focused on as part of your project (e.g., data quality management, prevention of data losses and leakage, management of the data lifecycle, safeguarding against potential bias subsumed in socio-technical systems and limiting the decision power of such systems to ensure fairness, accountability, and transparency, etc.)

4. Deliverables

Please provide an electronic copy of your final project submission as one zip archive by sending it to the course grader and instructor by the final project submission deadline as noted at the beginning of this document. The archive should include your project report and any other relevant details (in Microsoft Word format). The software portion of your project submission must be made available via a GitHub link and should include final versions of the various bits that were developed throughout the project (i.e., parts 1 through 4). The GitHub link should be clearly indicated as part of your report. You should name your archive using the following convention: lastname1_lastname2_final-project_su20.zip (note: this sample archive name provided assumes a team of two for illustration purpose).

5. Grading

[2] https://towardsdatascience.com/rootstrap-dikw-model-32cef9ae6dfb

The final project (part 4) will be graded on a maximum scale of 100 points and will serve as a final exam grade for the course. Your grade will be based equally on:

a. The overall quality of your documentation and project report.
b. The understanding and appropriate use of end-to-end database management systems and related technologies.
c. The quality and working status of your solution implementation.
d. Your ability to submit well documented solutions.
e. Extra credit may be granted for solutions that are particularly creative.

6. Additional Information

If you have not already done so, please let the course grader know as soon as possible about teaming arrangements (only two people per team). You will need to stay with the same team for the duration of the course. You should only submit one report/archive per team for each part of the project. To balance things out, the final grading for the course project will take into account the fact that you are working as a team instead of individually, so you should feel free to work individually as well.