New York University
Computer Science Department
Courant Institute of Mathematical Sciences

Database Systems Project Part III
Logical Schema Optimization and Machine Learning Model Creation

Due Date: 12/08/22

Course Title: Database Systems **Course Number:** CSCI-GA.2433-001
Instructor: Jean-Claude Franchitti **Session:** 8

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 insure 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 that improve user experience and organizational 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 is able to capture insights and drive decisions.

This next portion of the project focuses 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 focuses on the creation of a machine learning model to perform analytics on the unstructured data collected in the second part of the project. The end goal is to be able to drive business decisions that improve user experience and organizational excellence

Another project goal is to keep refining the "end-to-end" reference architecture that was developed in the second part 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. EDA Physical Database Design

The following steps should be followed to optimize the logical (RDBMS) schema and create a physical database design:

1. Perform and document physical database design for the database system and infrastructure selected in Part 2 of your course project. Techniques to be considered as part of your physical database design may include indexing, partitioning, clustering, and selective materialization as applicable. Please explain all choices being made and deploy your resulting design on a mainstream relational database system of your choice and preferably on one of the big clouds (e.g., Microsoft Azure). It is also fine to combine the use of local and cloud-based database management capabilities.

2. Identify appropriate business use cases to support a workflow-based application that enables a customer to obtain an insurance quote and a policy. Document the corresponding business use cases and the processes used by your application using a modeling notation of your choice. You do not need to implement a database program at this stage.

3. Machine Learning Model Creation

The following steps should be followed to create a machine learning model that performs analytics on the unstructured data collected in the second part of the project:

1. Refine and complement the use cases identified in the second part of the project. An example of a use case may be to leverage a dataset to forecast chronic diseases based on various

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

factors as explained in the project specification for part 2 of the project. Further establish your approach to evaluate Big Data ideas and consider ideas with short, medium, and long term benefits.

2. Select and train machine learning algorithm(s) to extract insights to help drive business decisions (e.g., creation of new insurance product, rate revisions, assessment of risk associated with a given insurance product, analysis of risks associated with the investment of insurance proceeds to generate additional revenue). The end goal is to be able to drive business decisions that improve user experience and organizational excellence on various fronts. You should feel free to select supervised (e.g., regression model, decision tree, Bayesian Networks, KNN, SVM, ANN), unsupervised (e.g., K-means clustering, DBSCAN, PCA, ICA, Anomaly Detection), and other learning algorithms (e.g., Reinforcement Learning, Deep Learning, Transfer Learning) as appropriate to meet your goals.

3. Keep developing your Big Data platform and data lake to compile various relevant datasets and further identify how insights that may be extracted from the data collected will feed into the EDA that was designed in the first part of the project. Your Big Data platform should be designed with growth in mind as you should be able to continually improve data analytics and visualization capabilities.

4. Elaborate further (based on your Big Data ideas and datasets) on the reference architecture that is most suitable for the insurance company to use in order to leverage hybrid data as part of their business.

5. Leverage Big Data Analytics and/or Streaming Big Data Analytics capabilities available via platform services provided on the big public clouds (e.g., Microsoft Azure Cloud) to extract, filter, store, analyze your datasets and present your analytics results.

4. Deliverables

Please provide an electronic copy of your homework submission as one zip archive by submitting it via Brightspace by the deadline as noted. The archive should include your physical model, machine learning model(s), and other relevant details along with your homework report (in word or text format). You should name your archive using the following convention for the homework archives: lastname1_lastname2_p3_fa22.zip (note: example project name assumes a team of two for illustration purpose).

5. Grading

All project assignments are graded on a maximum scale of 100 points. Your grade will be based equally on:

a. The overall quality of your documentation.
b. The understanding and appropriate use of end-to-end database management systems and related technologies.
c. Your ability to submit well documented solutions.
d. 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.