New York University
Computer Science Department
Courant Institute of Mathematical Sciences

Database Systems Project Part II
Logical Schema Optimization and Unstructured Data Collection

Due Date: 11/24/22

Course Title: Database Systems **Course Number:** CSCI-GA.2433-001
Instructor: Jean-Claude Franchitti **Session:** 7

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

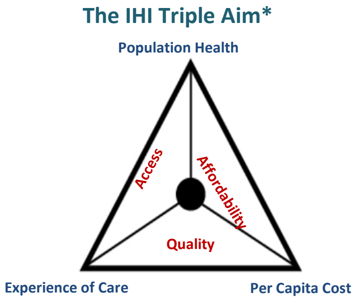
The first part of the project analyzed an existing logical model and led to the creation of a documented entity-relationship diagram using a mainstream software tool. The resulting model was partially validated against a set of business requirements and rules that were amended as/if needed.

This next portion of the project focuses on the collection of unstructured data to drive business decisions that improve user experience and organizational excellence. This should result 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 should inter-relate structured and unstructured data and result in an hybrid logical data model/data lake that can capture insights and drive decisions.

Another project goal is to put in place an "end-to-end" reference architecture that spans across the business, application, pyramid of knowledge (DIKW1), 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. Unstructured Data Collection

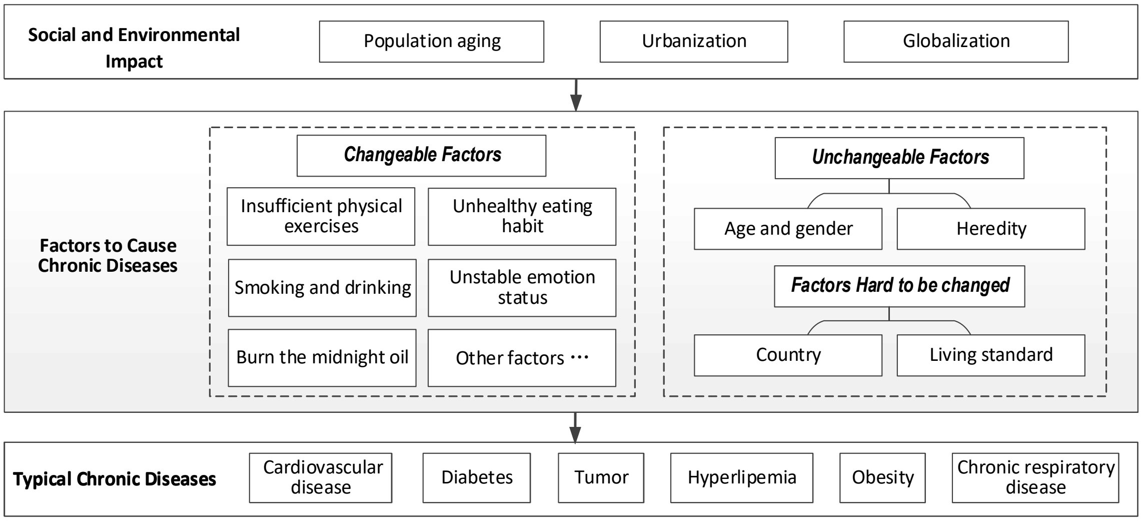
According to a recent Economics Intelligence Unit Report, the global yearly healthcare spend (HCS) per-person varies widely from $11,674 (US) to $54 (Pakistan), with a CAGR of 5.4%… Yet[2], the US ranks 54th globally in efficiency and 25th in life expectancy… As a result[3], the US healthcare spend is around 20% of GDP today. To fix the healthcare "burning platform", providers, patients and payers (i.e., insurance companies) should jointly achieve and sustain the Institute for Healthcare Improvement's (IHI) Triple Aim objectives illustrated in the Triple Aim paradigm shown in the diagram below.



Unfortunately, the traditional healthcare model requires a large number of patient visits with providers (i.e., accountable care organizations) that sometimes result in re-admissions at added costs for payer organizations. Therefore, while solutions are being developed to achieve the triple aim objectives, insurance companies still rely on health statistics collected by various sources (e.g., CDC) to forecast claims, create and refine insurance products, and update the corresponding products' rate books.

[1] https://towardsdatascience.com/rootstrap-dikw-model-32cef9ae6dfb
[2] Bloomberg 2018 Economies With the Most (and Least) Efficient Health Care Study
[3] National Healthcare Expenditure Projections, 2010-2020, CMS Office of the Actuary

The diagram below illustrates the factors that affect the detection accuracy of chronic diseases.



Your main goal in this part of the project is to collect datasets that could be used by the insurance company at hand to help forecast chronic diseases based on various factors, such as the ones illustrated in the diagram provided above. More specifically, you need to:

1. Put together a small data lake to compile various relevant datasets and 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.

2. Create/generate and optimize a logical database schema for the structured data used by the insurance company at hand based on the conceptual model created earlier. The logical database schema should inter-relate structured and unstructured data and support the hybrid data model/data lake required for the business to capture insights and drive decisions. Please note that you should not focus on generating actual insights in this part of the project. Please follow the steps in section 3 below (i.e., EDA Logical Schema Optimization) for details on how to create/generate and optimize your logical schema.

3. Elaborate 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.

4. Leverage a cloud platform (e.g., Microsoft Azure Cloud) to store and manage (all or part of) your hybrid data.

3. EDA Logical Schema Optimization

The following steps should be followed to create/generate and optimize your logical (RDBMS) schema:

1. Create and/or generate a logical schema that corresponds to the entity-relationship conceptual model developed earlier in the first part of the project. Please note that the tool you used to create the conceptual model may provide support to facilitate the generation of a logical schema for a database system of your choice. Please make sure that you select the database

system target that corresponds to the database product you plan to use to manage and store data as part of your project solution.

2. Optimize your logical schema and provide details as to the reasoning behind each one of the optimization steps you are taking. Please note that optimization in this context includes normalization as well as extensions required to inter-related structured and unstructured data.

4. Deliverables

Please provide an electronic copy of your homework submission as one zip archive by sending it to the course grader by the assignment deadline as noted. The archive should include your logical model and data lake design/implementation 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_p2_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 database 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.