New York University
Computer Science Department
Courant Institute of Mathematical Sciences

Database Systems Project Part I – Modeling

Due Date: 11/10/22

Course Title: Database Systems	**Course Number:** CSCI-GA.2433-001
Instructor: Jean-Claude Franchitti	**Session:** 6

1. Ongoing Project Background

Most enterprises today still rely on structured data stored in traditional relational databases and data warehouses. In addition to these data sources, enterprises are designing and deploying new data sources to manage large amounts of unstructured data, enable semi real-time big data analytics, and create machine/deep learning and/or AI digital solutions. To ensure growth companies have to put in place a reference architecture that spans across the business, application, pyramid of knowledge (DIKW[1]), and infrastructure domains. 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 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.

The first part of the project focuses on data stored in traditional relational databases and the use of such data 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 typically 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

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

- Operational Data Store (ODS) – Creation of physical database(s) that conform to the model.
- Roadmap – Means to move applications / operations to integrate with the ODS

The motivation that drives businesses towards adopting an EDA approach results from a myriad of data-related problems within the enterprise, including:

- Disjointed / flawed data sources
- Disjointed knowledge about data (experts in a single area / aspect, etc.)
- Too much data (no clear means of moving data as value changes over time, legal aspects, etc.)
- Crisis of data confidence (dirty data, difficult to uniquely ID customers / policies / associates, etc.)

As a result the architectural objectives for EDA include the need to establish the following:

- Data Retention Standards – manages data as its value changes over time (uniform migration / removal data aging data, consensus driven, in process with IT officers, etc.)
- Enterprise Data Model - a diagram (blueprint) that will support the design of storage systems (single integrated source for corporate data, consensus driven, etc.)

The following data storage terminology should be clearly understood when working with EDA:

- TDS (Transactional Data Store)
    - A simple, unambiguous means to track a company’s business
    - Actively being worked on by customers, customer service groups, etc.
    - EDM (long-term future state)
- ODS (Operational Data Store)
    - Daily snapshot of TDS to drive reporting, analytics, etc.
    - EDM (near-term future state)
- DW (Data Warehouse)
    - Tracks data that changes over time
- DM (Data Mart)
    - Stores aggregated and derived ODS / DW data for reporting purposes
- ETL (Extract Transform Load)
    - Provide opportunities for data cleansing

From a cost-benefit analysis standpoint, the following benefits can be attained via EDA:

- Single Source of Data
    - Increased consistency of reports, analytics, etc.
    - Reduction in Nightly Batch
    - Easier to affect changes (living model)

- Single Source of Knowledge about Data
    - Increased accuracy of reports, analytics, etc. (Data Quality Circles)
    - Easier to implement retention policies

2. EDA Modeling Questions

Refer to the sample PowerPoint slides provided for this assignment under demo programs on the course Web site (i.e., “Project Support Material).

1. Analyze the sample blueprint and create a corresponding entity-relationship diagram with full documentation. Please use Erwin Data Modeler (erwin.com) or an equivalent tool of your choice to draw the entity-relationship diagram.

2. Verify that the various cases described in the PowerPoint slides are supported by the model design and if not amend the design accordingly.

3. Deliverables

Please provide an electronic copy of your assignment submission as one zip archive by uploading it to NYU Classes by the assignment deadline as noted. The archive should include your E-R model and your assignment report (in word or text format). You should name your archive using the following convention for the assignment archives: lastname1_lastname2_p1_su22.zip. You are also required to provide a hard copy of your assignment report at the beginning of the class session on the date the assignment is due.

4. 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 related technologies.
c. Your ability to submit well documented solutions.
d. Extra credit may be granted for solutions that are particularly creative.

5. Additional Information

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 assignment. 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.