Systems And Methods For Big And Unstructured Data

Exercises

Christian Rossi

Academic Year 2024-2025

Abstract

The course is structured around three main parts. The first part focuses on approaches to Big Data management, addressing various challenges and dimensions associated with it. Key topics include the data engineering and data science pipeline, enterprise-scale data management, and the trade-offs between scalability, persistency, and volatility. It also covers issues related to cross-source data integration, the implications of the CAP theorem, the evolution of transactional properties from ACID to BASE, as well as data sharding, replication, and cloud-based scalable data processing.

The second part delves into systems and models for handling Big and unstructured data. It examines different types of databases, such as graph, semantic, columnar, document-oriented, key-value, and IR-based databases. Each type is analyzed across five dimensions: data model, query languages (declarative vs. imperative), data distribution, non-functional aspects, and architectural solutions.

The final part explores methods for designing applications that utilize unstructured data. It covers modeling languages and methodologies within the data engineering pipeline, along with schema-less, implicit-schema, and schema-on-read approaches to application design.

Contents

CHAPTER 1

Entity relationship diagram

1.1 Exercise one

Design an ER Model for a car rental system that manages the customers, the cars and their related elements. If a customer wants to rent a car, they should provide their personal ID card (9-digit ID, name,surname, birth date, address, release date, expiration date), driving license (10-digit drive licensenumber, name, surname, birthdate, release date, expiration date and the list of vehicles they are allowed to drive, e.g. truck, car, motorcycle, bus, etc.) and credit card (16 digit number and expiration date). Then, the list of cars is shown to the customer. Each car is described by its brand, the price per day, the maximum speed, and the fuel consumption per km. Before proposing the car, the system checks whether a car inspection has been performed in the last 3 months. Such data includes all the inspections (described by date and the name of the company who took care of it), alongside the list of all the operations the car underwent in each inspection. As soon as the customer picks the car they want, the keys are provided to the customer and the system stores the rental invoice (described by rental date, the period for which the car has been rented, and the final price). When the car is returned, the customer is charged. As soon as the payment is completed, the system marks the rental invoice as closed and the customer is provided with his invoice.

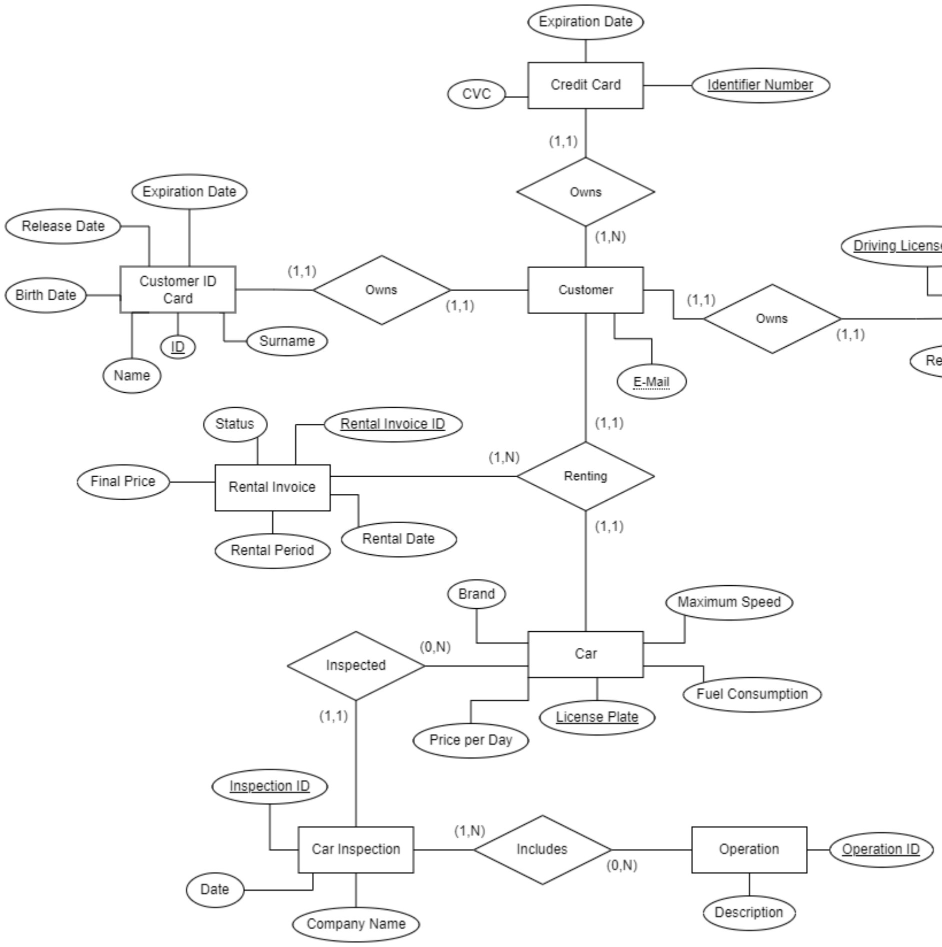
Solution

We have the following entites (with the corresponding attributes):

- CUSTOMER ID CARD (ID, Name, Surname, BirthDate, Address, ReleaseDate, ExpirationDate).
- DRIVING LICENSE (DriveLicenseNumber, Name, Surname, BirthDate, ReleaseDate, ExpirationDate, DriveCars, DriveBus).
- CREDIT CARD (IdentifierNumber, CVC, ExpirationDate).
- CAR (Brand, PricePerDay, MaximumSpeed, FuelConsumption, LicensePlate).
- CAR INSPECTION (Date, CompanyName, InspectionID).
- OPERATIONS (OperationID, Description).

- RENTAL INVOICE (RentalInvoiceID, RentalDate, RentalPeriod, FinalPrice, Status).
- CUSTOMER (E-mail).

To do so we assumed that the Customer is identified through an E-Mail. The diagrams of the given problem are:



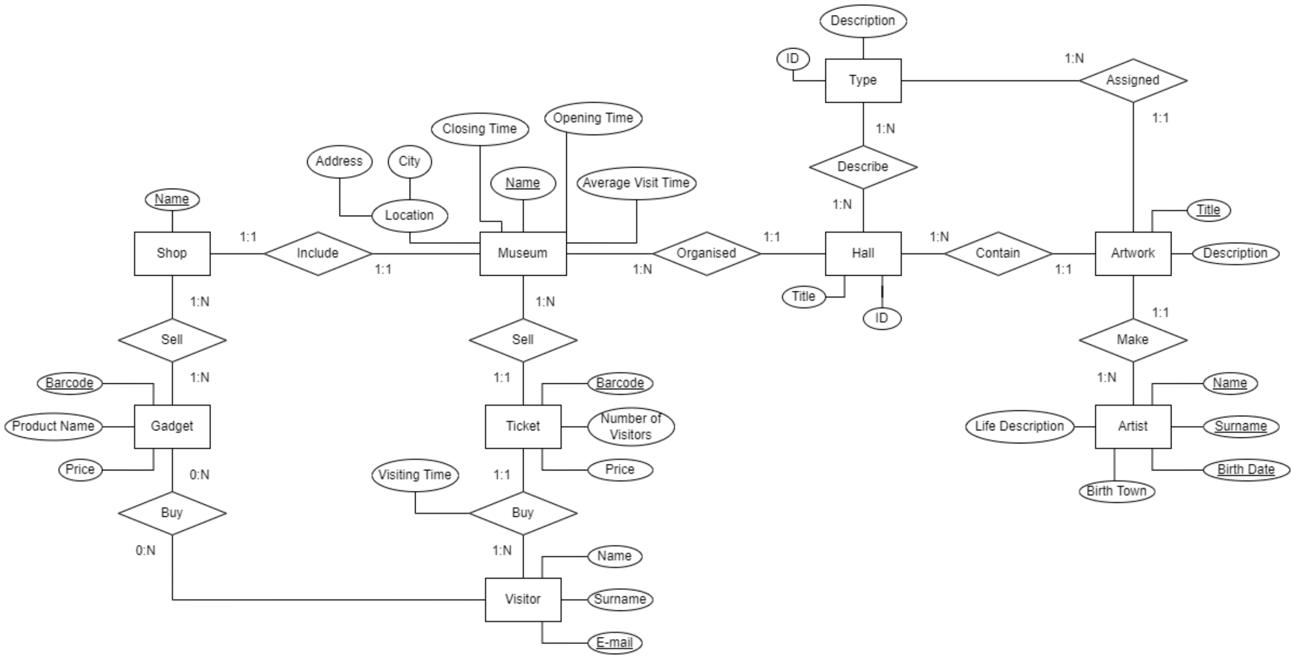
1.2 Exercise two

The Dutch Museum Association wants to keep records of all the Dutch museums and their artworks, as well as keep track of daily visitors and sales. Each museum is described by a unique name, location (further described by address and city), opening time, closing time, and the average time it takes to visit the whole museum. Opening and closing times are the same every day of the week. Every museum is divided into halls, and every hall contains one or more artworks. Each hall is described by a unique three-character identifier, title, and one or more types(e.g., sculpture, paintings, etc.). Artworks are defined by a unique title, description, and type (same as the hall types). Each artwork is produced by only one artist. Artists are described by name, surname, date of birth, town of birth, and their life description. Visitors buy tickets to visit the museum. Each ticket is described by a unique barcode, the number of visitors, and the total price. Visitors state the time when they would visit the museum when they buy tickets to allow the museum to keep track of the crowd. Museums also include a shop that sells gadgets. Each gadget is described by a unique barcode, product name, and price. Visitors buy gadgets at shops.

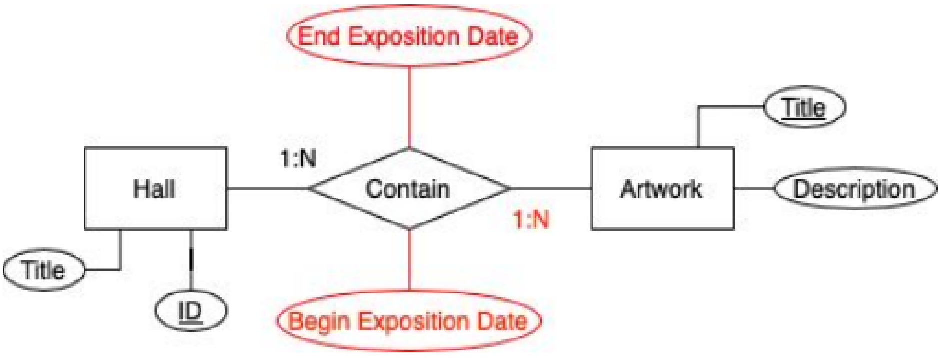
1. The Dutch Museum Association now wants to allow museums to borrow and lend artworks from other museums. They want to keep track of the period each artwork spends in each museum. The latter is agreed on before moving the artwork from one museum to another.
2. The Dutch Museum Association now wants museums to expand their shops. In particular, now museums will include three different kinds of shops: restaurant, sweets shop, and gadgets shop. While the first two shops sell food, the latter sells gadgets. Sweet shops only sell sweets. Food is described by the same attributes as gadgets, plus an expiry date.
3. The Dutch Museum Association wants to model the friendship relationship between the various artists whose artwork are exposed in their museum, considering sculptors and painters only. Furthermore, they made an astounding discovery: painters only befriend other painters and sculptors only befriend other sculptors.

Solution

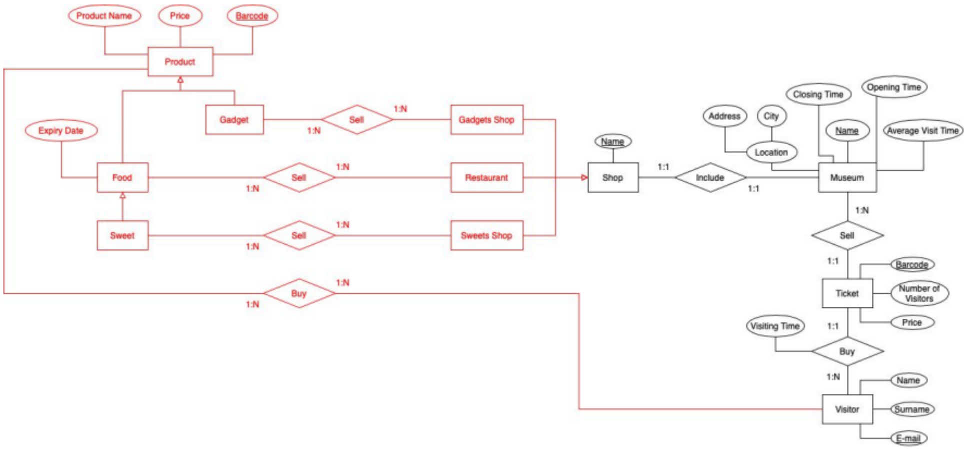
The corresponding diagram for the base problem is:



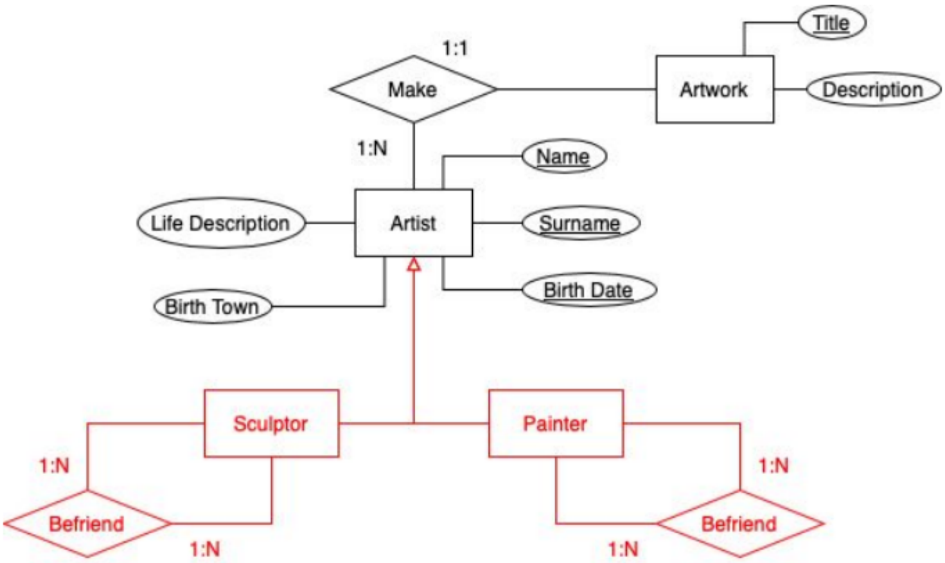
1. The added part is:



2. The added part is:



3. The added part is:


Title
1:1
Make
Artwork
Description
1:N
Name
Life Description
Artist
Surname
Birth Date
Birth Town
Sculptor
Painter
1:N
1:N
Befriend
Befriend
1:N
1:N