CSDS 234: Structured and Unstructured Data Spring 2026

Assignment 1: Relational Data Model

©**Instructor**: Prof. Yinghui Wu. **TA**: Yuhong Lu. @ CWRU **Due**: 29 Jan 2026 11:59 PM

Task 1 (30 points): Data Models and Types of Data

We have seen that the same dataset may be modeled differently depending on the application context, and that modern data systems need to deal with the challenges of *Volume, Velocity, Variety, and Veracity* (the "4Vs" as data challenges). For **each** of the following datasets:

- **(a)** Identify applicable data types, choosing from:
  - structured / semi-structured / unstructured
  - ordered data / time-series data
- **(b)** Identify one or two dominant "V(s)" among the 4Vs that you believe would be the major data challenge(s) to deal with.
- **(c)** Briefly explain your choices in (a) and (b). A dataset may belong to multiple categories, depending on the context.

Datasets

1. Smart home IoT data collected from thousands of households, including temperature, motion sensors, and power usage, streamed every second.

   Data Types: Structured data, Time-series data

   Dominant V's: Velocity, Volume

   Data is structured as the data collected such as temperature motion sensors and power usage typically follow a fixed schema and are typically stored in tables with the time-series since this is every second. The data is time-series as the data is generated continuously and indexed by time every second. The thousands of households creates massive amounts of data coming into the system every second making storage and retention a challenge and since the data is coming in at high speeds, the system requires fast ingestion and real time processing making velocity a challenge.

2. A large-scale social network dataset capturing users, friendships, and interactions (likes, comments, shares),used for community detection and recommendation.

   Data Types: Semi-structed data, unstructured data, ordered data

   Dominant V's: Volume, Variety

   Data is semi-structured as profiles, friendship graphs and interaction logs follow flexible schemas with evolving fields. Unstructured data such as comments, likes, shares and posts do not conform to a strict schema. Ordered data as the sequence can matter for interactions on the network for

modeling and recommendations. The large-scale social network makes volume a challenge as millions of users are generating a lot of data with all of the interactions which could make storage difficult. Since the dataset includes data such as profiles, graphs, posts, and shares, variety could be a challenge to deal with as there are many possible ways for the data to be collected.

3. Logs of user interactions on an online learning platform, recording page views, quiz attempts, and timestamps for millions of students.

   Data types: Structured Data, Time-Series/Ordered data

   Dominant V's: Volume, Velocity

   Data is structured as interaction logs should have well defined schemas making them suitable for log based storage of the students. Events are recorded with timestamps which makes the dataset time series and the order is important as since it is a learning platform, the order in which the student completes the interactions should tell whether or not they are successful. Since there are millions of students, volume could be a challenge to deal with since there is some much data, storing could be an issue with querying the data. Since the data should be generated in near real time, velocity could be a challenge are user interactions should continuously be happening as the online platform should be adaptive to help the student learn

4. A corpus of research papers (PDFs) used to build a semantic search engine for scientific literature.

   Data types: Unstructured Data, Semi-structured data

   Dominant V's: Variety, Volume

   Data is unstructured as pdf formatted papers are made of free text, images, figures, tables, and equations which do not follow a strict schema. Data is semi-structured as if the data is processed into specific areas based on titles, and references, the data can be structured to support indexing and search. Since papers can come from all different types of authors and publishers, the different layouts, writing styles and file formats could make extraction and normalizing a challenge. Since there is a large number of research papers, volume could be a challenge when trying to store indexes and embedding.

5. Autonomous driving data consisting of synchronized camera images, LiDAR scans, and GPS signals collected during road tests.

   Data types: Unstructured data, semi-structured data, time-series/ordered data

   Dominant V's: Variety, Volume

   Data is unstructured as images and Lidar scans are data types that can be collected without a fixed tabular structure. Data is semi-structured as the GPS signals follow defined formats that may vary across the different devices and configurations that are used. Data is time-series and ordered data as the sensor streams should be timestamped and synchronized and the order is critical for the perception and localization of the auto driving car. Since the dataset is combining many different data formats such as images, scans and GPS signals, variety could be a challenge in combining all the different datatypes. Since the autonomous vehicles should be sending in data as

much as possible for safety issues, volume can be a challenge in storing all of the different data.

6. A dataset of extracted features from medical images, represented as records with attributes {patient id, image id, feature vector, diagnosis label}.

   Data types: Structured data, ordered data
   Dominant V's: Veracity and Volume

   Data is structured as the fixed schema is given with attributes such as patient id, image id, etc. which makes it suitable for relational databases. The data is ordered as if the images are collected over time, the records could potentially be ordered by the scan date in which they were scanned by the nurse/doctor but may not be important for the collection. Since the data is medical and having accurate data is critical, veracity could be a challenge as errors in the extraction or misdiagnoses can directly affect the models reliability. Since there could potentially be a lot of images, the database may have a large number of patients needing a substantial amount of storage, leading to volume being a challenge.

7. Product listings from multiple e-commerce websites, collected via web scraping and stored in JSON format.

   Data types: Semi-structured data, unstructured data

   Dominant V's: Variety and Veracity

Data is semi-structured as the file is in JSON format which provides a flexible schema with key values and nested objects. Data could also be partially unstructured as fields such as product descriptions, reviews or specification are often free text and do not follow a structure. Since the different e-commerce sites use different layouts, names for the descriptions of sorting, categories and levels of detail, variety can be a challenge with the inconsistency between all of the different websites. Since the internet can be a little biased with potential bots and with noise from biased review, veracity could potentially be a challenge with other potential human error like outdated listing and proxies.

8. Real-time clickstream data from a large online advertising platform used to update bidding strategies.

   Data types: Structured data, Time-Series Data

   Dominant V's: Velocity, Volume

Data is structured as clickstream data can contain the elements for the user that is interacting such and be put into a log which can follow a schema. Data is time-series as the events arrive continuously, it is very likely that they are indexed by the time in which the click occurred, where the event order is important since the company is tracking bidding strategies which need to track when the clicks occurred. Since the clickstream data is most likely coming in very fast, the data must proceed in real time to update the bidding strategy, making velocity a possible challenge with the dataset. Since advertising is marketed towards millions of people, the volume of people clicking on the bidding strategies will likely cause a lot of volume and need a lot of storage, making volume a potential challenge.

Task 2 (30 points): Relational Data Model & Integrity Constraints

Consider the following schema for **FreshDash**, an online food delivery and recommendation plat- form. It keeps track of restaurants, customers, menu items, orders, deliveries, and customer reviews. An order represents a purchase placed by a customer from a restaurant at a specific time.

Schema (6 relations)

Restaurant(rest_id, rest_name, city, street_address, cuisine_type)

MenuItem(item_id, rest_id, item_name, category, price)

Customer(cust_id, cust_name, email, phone, home_zip)

Order(order_id, cust_id, rest_id, order_time, total_amount)

OrderItem(order_id, item_id, quantity, item_price_at_order_time)

Review(review_id, order_id, cust_id, rating, review_time)

Validated facts

The following facts have been validated.

- Each restaurant has a unique rest id; the combination (rest name, street address, city) is also unique.
- A restaurant can offer many menu items; each menu item belongs to exactly one restaurant.
- Each customer has a unique cust id and a unique email.
- A customer can place many orders; each order is placed by exactly one customer.
- Each order is associated with exactly one restaurant.
- An order can include multiple menu items, and a menu item can appear in many orders.
- For each (order id, item id) pair, there is at most one OrderItem record.
- A customer may write at most one review per order.
- Ratings are integers in $\{1, 2, 3, 4, 5\}$.

Questions

- **(10 points)** Identify a primary key and a candidate key for each relation.

  Restaurant

  Primary Key: rest_id

  Candidate Key: rest_name, street_address, city)

  MenuItem:

  Primary Key: item_id

  Candidate Key: item_id

Customer:

Primary Key: cust_id

Candidate Key: email

Order:

Primary Key: order_id

Candidate Key: order_id

OrderItem:

Primary Key: (order_id, item_id)

Candidate Key: (order_id, item_id)

Review:

Primary Key: review_id

Candidate Key: (order_id, cust_id)

- **(10 points)** Identify the foreign keys of each relation, given your choice of primary keys.

  Restaurant: None

  MenuItem: rest_id from restaurant

  Customer: none

  Order: cust_id, rest_id from customer and restaurant

  OrderItem: order_id, item_id from order and item

  Review: order_id, cust_id from order and customer

- **(10 points)** Identify the foreign keys that are also a part of the primary keys of the same relation they are defined on.

  For OrderItem order_id and item_id are the primary keys for Order and MenuItem respectively. Since an order item is unique based on the order and what item is on the menu, the combination is the primary key for the OrderItem class. The other foreign keys identified above are primary keys, but are not primary keys for the same relation that they are defined on.

Task 3 (20 points): Schema & Data Constraints

Continue with the above schema. FreshDash wants to support **personalized food recommendations**. To do this, the system must track:

- which customers ordered which menu items,
- from which restaurants,
- and at what time.

Questions

- **(10 points)(a)** Based on the existing schema: identify the relation(s) needed to answer these questions.

  To track which customer ordered which menu item from which restaurant and at what time, the relation of order will be needed as in the order relation, the customer, restaurant, order time which are all needed for the personalized food recommendations. The OrderItem relation will also be needed to identify what was the menu item associated with the order. The path of the relationship would start as a customer, which has a foreign key in order which is used to track the customer, the restaurant and the time and then to the OrderItem which will associate the order with the menu item that was purchased.

- **(10 points)(b)** As a data model designer, you need to monitor potential "harmful" data modifications that may cause violations. Choose any three data constraints on the FreshDash schema, and for each data constraint, give one possible modification that may cause the violation of the constraint. Briefly explain — you can give an example with some tuples.

  Three data constraints on the FreshData schema can be entity integrity constraints, referential integrity constraints and domain integrity constraints.

  Integrity constraints which is the primary key constraint can be exemplified with the constraint that every customer in the customer table must be unique and cannot be null as there cant be two of the customer and every order must come from a customer. An insert attempting to add a new customer to an id that already exists is a harmful modification. An example could be that Alice has the cust_id of 10, if Bob was attempted to be inserted with a cust_id of 10, a violation should trigger preventing that from happening.

  Referential Integrity constraints which is the foreign key constraint can be exemplified with an item in the orderitem table must be found in the menuitem table. A harmful modification can be a delete operation on the menuitem table that could remove a menuitem from the ordered table. An example of this is if a restaurant offered Lasagna and referenced it as item_id of 52 and there are existing records of item_id 52 in the OrderItem relation, deleting the Lasagna would leave the records orphaned and the system would not know what the customer ate.

  Domain integrity constraints can be exemplified with a review being between the integer range of 1 to 5. A harmful modification can be viewed with the insert operation of a value that is out of range. An example of this is if the customer tries to give a review of -1 or 6 and the system accepts it, the averages of the restaurant's scalability will skew the averages of the restaurant's review which could lead to inaccurate recommendations.

Task 4 (20 points): R&D Practice

Get familiar with at least one relational data system (e.g., MySQL or PostgreSQL).

- Download and install a system of your choice. We provided an installation guide for Post- greSQL in Module 2. Here are useful links if you choose MySQL:
  - https://dev.mysql.com/downloads/installer/
  - https://dev.mysql.com/doc/mysql-installation-excerpt/5.7/en/
  - Play with MySQL tutorial: https://dev.mysql.com/doc/refman/8.0/en/tutorial.html
- Build your own toy structured dataset with the “Healthsense” toy dataset, provided as a running demo in Module 2.
- Set up a Github repo to manage your R&D project and relevant code for CSDS 234.

For your answer

Submit the following (either hosted in your GitHub project, or uploaded in a zip folder):

- **One SQL file**, named first name 234 A1.sql. This file must include:
  - CREATE TABLE statements
  - Primary keys
  - At least one foreign key
  - At least one constraint (CHECK or NOT NULL)
  - INSERT statements (3–5 rows per table)
- **Supplement materials** with prefix ”first name 234 A1”:

  run ”Select * from ...” statements, to show that you have successfully loaded the csv to the structured relations. Take screenshots of the results and submit them.

  (Optional, no need to submit): try some “harmful” insertion that may violate the integrity constraints. What will PostgreSQL report? Observe and think about why.

  Here is a link to my GitHub: https://github.com/pjm187/CSDS234Assignment1

How to submit your work

Please submit your work via Canvas. By default, all submissions should be **typed**. If it has to be made handwritten, the handwriting must be clear, neat, and fully legible. Illegible or unrecognizable handwriting may result in loss of credit, as the work cannot be properly evaluated.