**Relational schema design exercise**

**(Submitted by Ravikanth Jonnalagadda – responses are in blue text)**

**Overview**

The setting is an auto dealer. In this company, there are 3 departments, including:

* Inventory
* Sales
* Customer relations

Right now, each department manages their information differently.

These departments would like to integrate their data into a shared database, in order to be able to answer questions like, “What engine is in Customer Smith’s car?”. It’s a challenge to answer right now because the Sales department has information about which car Customer Smith bought; the Inventory department tracks which engines are in which cars; and the Customer Relations has totally separate, slightly redundant information about Customer Smith.

You have been asked to design and populate a database that is effective and efficient for all departments. You are given data from each department:

* File A (Inventory): Document relating models to styles, power trains, quantities, and individual cars, etc.
* File B (Sales): Document relating people to cars, prices, trade-ins, etc.
* File C (Customer relations): Document relating people to personal information, information about services, warranties, etc.

**Instructions**

1. Write a short narrative description of each file we have given you. What information do they contain? How is the information formatted and organized? What kinds of information is shared among the files? (50-100 words per file)

**MDS\_Exercise1\_FileA.txt**:

* This is a text file containing the Dealer’s Inventory information.
* Each row in the file contains Tab Delimited data and ends with CR (\r).
* Attribute names are not provided and need us to make assumptions about the details provided. This file’s intent appears to be to provide the following information - Inventory Item#, VIN, Year, Make, Model, Style, Drivetrain, Color, Type, Engine Type, Price
* Some rows have empty detail (eg: Style).
* Inventory file detail is being used in Sales File to get price detail
* Special care is being taken to provide Price in double quotes.

**MDS\_Exercise1\_FileB.csv**:

* This is a CSV (Comma Separated values) file containing the Dealer’s Sales information.
* Each row in the file contains comma delimited data and ends with CR (\r).
* Attribute names are provided in the first row
* This file repeats some of the Customer (eg: Address) and Vehicle information (eg: Make, Model, etc.) by including them with Sales Information.
* Some rows have some missing values
* For any application that processes this file – unless special attention is paid to the treatment of the Discount, Price, etc. detail – it will result in incorrect processing as the comma within will cause issues.

**MDS\_Exercise1\_FileC.docx**:

* This is a MS Word document file containing the Dealer’s Customer information.
* This file has data organized in blocks, with each block having 4-5 rows of detail.
* This has no header block describing the detail included in the document. The intent behind each block seems to be providing detail as follows
  + Row 1 : Customer’s Name
  + Row 2 : Customer’s Street Address
  + Row 3: Customer’s City, State, Country, ZIP
  + Row 4 : Customer’s Profession
  + Row 5: Customer’s Need to Finance
* This file format is very proprietary and applications should take special care when processing it.

2. Use your narrative to design a database schema that will accommodate the information in the files. Your schema should include:

* Tables
* Attributes (columns) in each table
* Datatypes for each attribute
* Primary and foreign keys

Workbook is uploaded as a separate artifact, and it addresses both Step 2 & Step 3.

3. Create an example of each table, populated with data from the files.

4. Write an account of your process for creating the database scheme and tables. Describe any decisions they had to make as you developed a relational schema.

The prior data model upon analysis was determined to be not conducive to support some of the basic tenets of Data Curation including preservation, discoverability, access, workflow, identification, etc. We can see some of the issues outlined in the short narrative in step(1)

* How did you decide to represent the data in the way that you did?

Given the state of disorganization in the prior model eg: disparate standards/practices, lack of model definition, siloed data, redundancy, etc. this effort needs to stream line the data model and come up with something that supports the basic tenets of Data Curation. The relational Model is a great way to accomplish this.

Towards this end, data across the various formats has been analyzed to see what model will best serve the interests of the organization and support preservation, discoverability, access, workflow, identification, etc.

Also, upon analysis of prior model – it was noticed that there is lot of duplication of data across the divisions and it became apparent that coming up with a relational model that is normalized will help ensure that updates can be kept to a minimum and the Schema will be better manageable, maintainable and efficient.

Towards this – it felt prudent to first identify the key relations (tables), their attributes (columns), the data type of the columns, the constraints that apply over these, and potential processing rules that will add benefit.

* Did you leave out any information? If so, why?

This data model takes into account all of the business needs that the prior data model was supporting, as well as the short-comings of the prior data model. So, a conscious decision was made not to leave out any information need addressed by the prior data model. The goal here was to ensure better support for business needs and adherence to the data curation principles so that the true value of data can be unlocked by the business.

However, given the scope of this task – a few things have been left out of the schema definition including data validation specifications, etc. Also, it was chosen to make missing values in the original data sources as null when populating them in the new relational model. Similarly, very few processing rules for data are specified. Again, keeping in mind that the focus is primary on creation of the Schema – not much information is provided on Schema versioning, metadata management, etc.

There are quite a few Categorical columns in the Schema, however the reference table related detail has not been provided in the response to keep the schema reasonable for the scope of this engagement. On a related note, the Audit table and its usage has been limited to inclusion for demonstrating how it can be used, and is by no means complete or exhaustive.

All data and information needs addressed by the prior model are also addressed by the new model, in addition to adhering to relational data model best practices.

* Why did you choose certain things as attributes? As keys?

Attributes were chosen based on their alignment with the Relation (i.e. table) being specified. If it made sense to follow rule of normalization and express some of the attributes through a different table, then those choices were made. The keys in the new model are primarily Table ID columns containing Integer values.

There could have been other choices made – particularly for the Inventory table, a VIN might have been a decent choice to add to the key, but it was decided to retain the Inventory ID as the key column – for a number of reasons including the VIN was not being validated (and all the VINs provided were not valid – ex: they contain invalid characters like I or have a length of more than 17, etc.).

In a similar fashion, for the Customer Table, it was decided to leave Address details (Street Address, City, State, Country, ZIP, etc.) as part of the customer table and not abstract it out to keep the model simple for the scope of this assignment.

* What were the hardest decisions you had to make in this design process?

The hardest decisions were centered around striking the right balance in integrating the data from the three divisions together to support the needs of the Dealer, and determining how much normalization will be appropriate to support the needs of the business without being a hindrance to usage, adoption, maintenance and support.

* How does your schema design support data independence?

“Abstraction implemented with indirection supports data independence.” Adhering to these principles, the schema design ensures that someone using this model at runtime (implemented through an RDMS) will be able to interact with the relational constructs and keep their interactions to the logical level without having to worry about the physical storage details.

Thus this schema design supports data independence. In the prior state (the three files), that was not the case - #1 there is no relational construct based interaction in reading from text or CSV files (although other methods of interaction exist), and the MS word file is a totally different beast altogether).

* How may your schema design support the overarching goals of data curation (revisit objectives and activities of Week 1)?

This schema design supports the various goals of data curation as follows:

* Provides an appropriate data model using appropriate standards to support **Organization**
* Supports reliable **Storage** through facilitating implementation of an RDMS
* Ensures **Preservation** through using Relational Model – it ensures data is understandable and useable in the future (compared to the current state as represented by the three files).
* Through specification of easy to follow schema and through other constructs supported by the relational model supports **Discoverability** of relevant data.
* This Schema provides for **Access** to retrieve and distribute data for us.
* This relational model can be deployed on an RDMS supporting the ability to identify, authenticate and validate data much better than the current state can support.
* Which curation activities could enhance or sustain the database for future discovery and use for new purposes? What additional activities would you recommend?

Additional recommended activities include :

* support for Provenance
* metadata management
* ability to master data
* validation of data with internal and/or external sources (eg: Vin check, Address check, etc.)
* (entity resolution) support to ensure ability to resolve conflict if same keys show up with differing attribute values during integration
* Ability to manage reference data, etc.