Assignment 1 – Relation Schema Design Exercise
- RAJA SHEKAR BOLLAM (bollamr2)

[1. 20 Points Narrative Description]

File A
This file has Inventory Information in this case a car. Data on inventory owned by a person seems to be uniquely identified by an id. This ID is known as VIN as referenced in File B.
It seems to contain information as follows - Year of Purchase of the car, Name of the Manufacturer, Model of the Car, All Wheel Driver or a 4 Wheel Drive, Color of the Car, Number of Car doors, Type of Engine and Price of the Car. Price of the Car is quoted as US Dollars and look consistently formatted in the data set. Data seems to be complete but not arranged in any specific format or Order.
In the context of Completeness, there don't seem to be any missing values or errors
As for the readability, data can be cleaned and arranged in a much better format.
By looking at the data, it is difficult to identify the person who owns a specific inventory as there is no information related to the person.

File B
This file seems to have a lot of sales information in regard with the person and his ownership of a car. It has information related to information on the person who owns a particular inventory and details on the inventory.
In the context of Completeness, there are several columns missing for some people.
In the context of Redundancy, some of the Inventory details are duplicated from File A.
Since this file is expected to have information related to Sales, data needs to be re-organized.

File C – Person Information
This file seems to contain Person Information of the Person who owns a particular inventory. Person Information includes Name with address, Occupation and some other notes.

[2. 10 Points Narrative to design a database schema]

Schema with Tables, Attributes, Data Types, Primary and Foreign Keys.

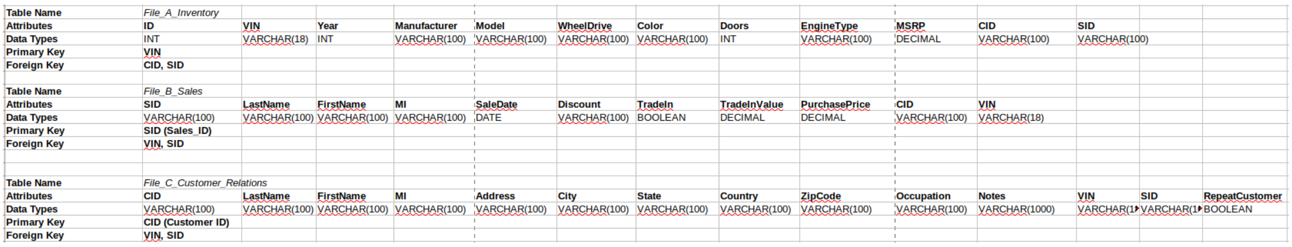
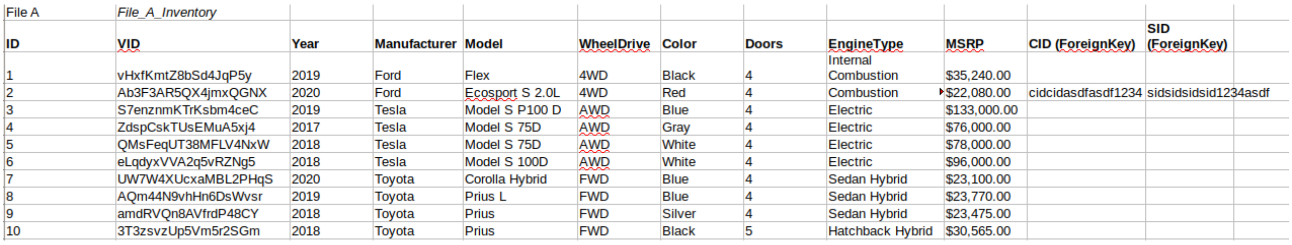


Table Name	*File_A_Inventory*													
Attributes	**ID**	**VIN**	**Year**	**Manufacturer**	**Model**	**WheelDrive**	**Color**	**Doors**	**EngineType**	**MSRP**	**CID**	**SID**		
Data Types	INT	VARCHAR(18)	INT	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	INT	VARCHAR(100)	DECIMAL	VARCHAR(100)	VARCHAR(100)		
Primary Key	**VIN**													
Foreign Key	**CID, SID**													
Table Name	*File_B_Sales*													
Attributes	**SID**	**LastName**	**FirstName**	**MI**	**SaleDate**	**Discount**	**TradeIn**	**TradeInValue**	**PurchasePrice**	**CID**	**VIN**			
Data Types	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	DATE	VARCHAR(100)	BOOLEAN	DECIMAL	DECIMAL	VARCHAR(100)	VARCHAR(18)			
Primary Key	**SID (Sales_ID)**													
Foreign Key	**VIN, SID**													
Table Name	*File_C_Customer_Relations*													
Attributes	**CID**	**LastName**	**FirstName**	**MI**	**Address**	**City**	**State**	**Country**	**ZipCode**	**Occupation**	**Notes**	**VIN**	**SID**	**RepeatCustomer**
Data Types	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(100)	VARCHAR(1000)	VARCHAR(1	VARCHAR(1	BOOLEAN
Primary Key	**CID (Customer ID)**													
Foreign Key	**VIN, SID**													

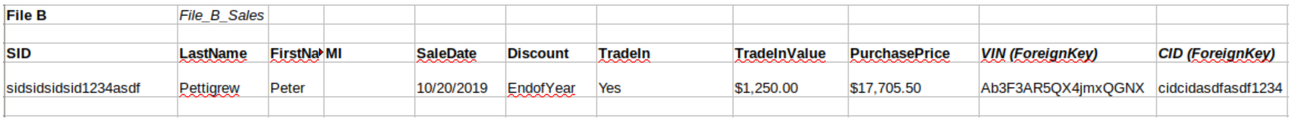
[3. 15 Points Example of each table]

File A - Inventory



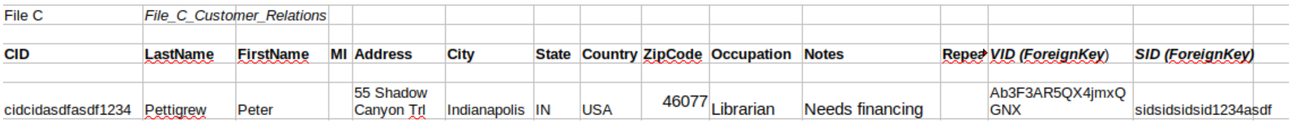
File A	*File_A_Inventory*											
ID	**VID**	**Year**	**Manufacturer**	**Model**	**WheelDrive**	**Color**	**Doors**	**EngineType**	**MSRP**	**CID (ForeignKey)**	**SID (ForeignKey)**	
1	vHxfKmtZ8bSd4JqP5y	2019	Ford	Flex	4WD	Black	4	Internal Combustion	$35,240.00			
2	Ab3F3AR5QX4jmxQGNX	2020	Ford	Ecosport S 2.0L	4WD	Red	4	Combustion	$22,080.00	cidcidasdfasdf1234	sidsidsidsid1234asdf	
3	S7enznmKTrKsbm4ceC	2019	Tesla	Model S P100 D	AWD	Blue	4	Electric	$133,000.00			
4	ZdspCskTUsEMuA5xj4	2017	Tesla	Model S 75D	AWD	Gray	4	Electric	$76,000.00			
5	QMsFeqUT38MFLV4NxW	2018	Tesla	Model S 75D	AWD	White	4	Electric	$78,000.00			
6	eLqdyxVVA2q5vRZNg5	2018	Tesla	Model S 100D	AWD	White	4	Electric	$96,000.00			
7	UW7W4XUcxaMBL2PHqS	2020	Toyota	Corolla Hybrid	FWD	Blue	4	Sedan Hybrid	$23,100.00			
8	AQm44N9vhHn6DsWvsr	2019	Toyota	Prius L	FWD	Blue	4	Sedan Hybrid	$23,770.00			
9	amdRVQn8AVfrdP48CY	2018	Toyota	Prius	FWD	Silver	4	Sedan Hybrid	$23,475.00			
10	3T3zsvzUp5Vm5r2SGm	2018	Toyota	Prius	FWD	Black	5	Hatchback Hybrid	$30,565.00			

File B – Sales



File B	*File_B_Sales*									
SID	**LastName**	**FirstNa**	**MI**	**SaleDate**	**Discount**	**TradeIn**	**TradeInValue**	**PurchasePrice**	***VIN (ForeignKey)***	***CID (ForeignKey)***
sidsidsidsid1234asdf	Pettigrew	Peter		10/20/2019	EndofYear	Yes	$1,250.00	$17,705.50	Ab3F3AR5QX4jmxQGNX	cidcidasdfasdf1234

File C – Customer Relationship



File C	*File_C_Customer_Relations*												
CID	**LastName**	**FirstName**	**MI**	**Address**	**City**	**State**	**Country**	**ZipCode**	**Occupation**	**Notes**	**Repea**	***VID (ForeignKey)***	***SID (ForeignKey)***
cidcidasdfasdf1234	Pettigrew	Peter		55 Shadow Canyon Trl	Indianapolis	IN	USA	46077	Librarian	Needs financing		Ab3F3AR5QX4jmxQGNX	sidsidsidsid1234asdf

[4. 30 Points Process for creating the database schema and tables. Describe decisions]

Relationship between data files
Upon careful observation on the data provided File A, File B, File C, we come down to the following conclusion on the relationship between these 3 data files.
File B has all the data related to a person’s information, Inventory (car) he possess and his purchase information o the inventory. A VIN column in File B is an Unique ID that can be used to identify the owner in File A. The First and Last Name in File B can used to get the address of the owner from File C.

Process Followed
Below is the Process followed for creating the database schema. We have divided entire data in to 3 tables. File_A_Inventory, File_B_Sales, File_C_Customer_Relations

File A - Inventory
1. This table needs to have information about the purchased car.
2. VIN is used as Unique Key/ID to identify a record.
3. Year, Manufacturer, Model, WheelDrive, Color, Doors, EngineType, MSRP are identified as the attributes related to a car. Hence they are group in this table.
4. We added foreign keys CID(Customer ID), SID(Sales_ID) to refer the customer and Sales information.

File B – Sale
1. This table needs to identify a person who needs to buy/bought a car.
2. The inventory details need to brought from File A if required, based on the created VIN number.
3. We need to have some person details related to the person and some sales related information on the purchased item.
4. We identified SaleData, Discount, TradeIn, TradeInValue, PurchasePrice, RepeatCustomer, VIN as information related to Sale of a Car.
5. VIN Is identified as foreign key that can be used to retrieve car information from File A table.
6. Last Name, First Name, MI are NOT used as a primary key for this table, since there could be two different people with same names and we cannot uniquely retrieve information on the name.
7. We added SID(Sales_ID) as a Unique identifier for a entry in the table. This is used as Primary Key for the table. Same key can be used as a Foreign key to File C and File A to retrieve customer related information.
8. We could as well move “Discount” to another table or create a new Table, but I choose to keep it in File B. In case in future a scenario arise where I need to have type of Discount, we could either create a new master table or move it to File C.

File C – Customer Relations
1. This table need to have information related to the person, and the kind of customer he is to the company.
2. As a result we have identified Address, City, State, Country, ZipCode, Occupation and Notes as the required attributes for this table.
3. Notes has customer specific information such as “Needs Loan”, “Inquiring about Car” etc.
4. Last Name, First Name, MI are NOT used as a primary key for this table, since there could be two different people with same names and we cannot uniquely retrieve information on the name.
5. We added CID(Customer_ID) as a Unique identifier for a entry in the table. This is used as Primary Key for the table. Same key can be used as a Foreign key to File B and File A to retrieve customer related information. We could use Customer_ID as an incremental integer as well. As a preference of choice, I choose it to be a VARCHAR(100).

6. We moved “RepeatCustomer” field from File B to File C as per the review comments, since it is better to get a feel of customer type, by knowing the customer id. It helps in giving some customers preferential treatment based on this attribute.

As for the Data Types for each attributes are concerned we have used “VARCHAR” with a length of 100 for all character related entries. INT for number related attributes, DECIMAL/BOOLEAN for items related to this kind of data and DATE data type for SaleDate.

Questions & Answers

- How did you decide to represent the data in the way that you did?
    1. The priority criteria employed in representing the data is data abstraction. Attributes in one table should be able to represent all the details of that type of Table. For example: Customer_Relations table should have all the information related to customer. That way, we don’t need to access multiple tables.
- Did you leave out any information? If so, why?
    1. No information is left out, but redundant attributes are removed and moved to tables that best describes those attributes.
    2. It is absolutely necessary, since read, write, update operations will be brought down to minimal number. It also helps in removed inconsistencies in the tables.
- Why did you choose certain things as attributes? As keys?
    1. The primary criteria of a Key is to uniquely identify table entries. Initially we thought of using “Firstname+MI+LastName” as primary key. But this combination cannot be unique. There is always a chance of two people having same combo of names.
    2. We have decided to use Uniquely generated keyword with a size of 100 Chars (size is a matter of choice. It could be modified based on requirements)
    3. VIN, CID(CustomerID), SID(SalesID) are using as primary keys for the tables.
- What were the hardest decisions you had to make in this design process?
    1. Placing which attribute in which table is one of the hardest decisions to make. It has impact on Data Consistency, Performance and Ease of Usage of tables to deliver better services to the customer.
- How does your schema design support data independence?
    1. To support data independence we paid extra attention to remove and duplicity of attributes in different tables.
- How may your schema design support the overarching goals of data curation (revisit objectives and activities of Week 1)?
    1. Data is organized efficiently in a way to support analysis of data and reuse of the tables over a period of time. We could add more attributes in the table or create more tables to support ever growing requirements.
    2. Activities like Collection, Organization, Storage, Preservation, Discover-ability, Access, Workflow, Identification, Integration, Reformatting, Reproducible, Sharing, Modification have been addressed.
    3. Activities such as Security, Compliance, Provenance, Communication have not been addressed in this assignment.
- What are the pros and cons of your schema design?

  Pros:

i. Data is well organized with attributes describing the necessary information that the table requires.
ii. Removed inter-dependency between the tables, by removing duplicity of attributes.
iii. Used uniquely identifiable keys as Primary Keys so that a query could return single entry from the table.
iv. Attributes in the table have well reasoned data types.

Cons:

i. The tables and attributes that list all the attributes that might be required in real life scenario to provide best customer relations treatment. We could add other attributes to the table for instance, we could add “DiscountType” attribute to get to know what kind of discount did a person get while buying a car.
ii. DataTypes for the attributes could be customized based on the requirements of usages of those attributes. For simplicity purpose I choose “VARCHAR” is most of the cases.
iii. We didn’t add NULL checks for certain attributes. It is suggested that we add NULL checks.

- Which curation activities could enhance or sustain the database for future discovery and use for new purposes? What additional activities would you recommend?

  i. Sales, Inventory, Customer_relations are too generic information for a Cars Sales details. We could add more tables to make the table schema more robust. Such as, Discount Information, Notes, Car Type and Engine Type into different tables.

  ii. We could add NULL Checks to attributes that need it.

  iii. We could customize the Data Types of the attributes.

  iv. We could add more tables to enhance data abstraction, performance and ease of usage of the tables.