**Overview of the Dataset & Goals**

The data set contains 3 different files about an Auto dealership containing sales, customer information, product inventory. Our objective is to create a relational schema set that can be used to answer various analytical questions.

**Data Analysis**

Each one of these files provides comprehensive information from their specific dimension (Sales, Inventory, Customer). Examining the data from all 3 files tells us that data can be corelated, but it has to be cleaned and curated in order to make it ready for the analysis. Few important keys that can be used for co-relation are:

* VIN number is a common key between sales and inventory files.
* Customer’s first Name, Last Name, address combination is used to create a unique customer key called CUSTOMERID. This can be used as a common key between sales and customer files.

**Data Quality**

For the source data presented in different files the following quality issues were noted.

* Data Format -The data format for all 3 files is not consistent. Inventory information is in a text with tab separated values for columns, Customer information is in a word doc in an unstructured format, Sales information was a csv file.
* Missing Information – One customer’s city information was missing, for one customer’s state information was missing and for some customer’s country information was missing
* Data capturing – Some data entry errors were noted like tab spaces were not consistently given between fields (in the inventory files). When Middle initial is not present appropriate space was not given. MSRP was missing for record #9 in sales file, Purchase Price is missing for record #3.
* Data Consistency – Some data is not consistently available across all 3 files. For example., customer’s zip code is not present in the sales file, but, is present in the customer file. Together both sales and customer files can be used to enrich a customer record to have a zip code.
* Data duplication – Customer data is duplicated between customer and sales data.
* Attribute classification – Sales file has column names that classify the records data, but, other files doesn’t have this classification. By looking at the data in Inventory and Customer we can identify these attributes like zip code , make , year column etc.,

**Data Cleansing and Curation**

To address the above-mentioned issues with the Data Quality, we need to have a transformation process that moves the data from flat files to a structure excel sheet where we can analyze and filter/correlate the data. This gives us data independence from physical schema.

To cleanse and curate the data, I’ve done the following:

* Converted and imported the data into excel sheets and separated each attribute into a column
* Manually corrected typos/tab errors
* Enriched customer information in sales sheet with the data in customer sheet and vice-versa.
* Enriched MSRP price in sales records with the MSRP from inventory data
* Added the appropriate attribute/column name to the inventory and customer sheets
* Created UniqueIds to track Customers

**Design process**

Most of the analytical needs are centered around sales transactions. For this reason, I’ve decided to organize the data around sales information. To organize the data around sales, star schema design fits well. Following this pattern, all the sales facts are tracked in the ‘FACT\_SALES’ table and other associated keys are tracked in different dimensional tables. All the dimension tables have a primary key and they were represented as foreign keys in the star schema tables. This design enables extensibility to dimension tables i.e. new fields can be added without touching other tables.

From the data reorganization perspective, none of the fields were actually dropped as part of this schema organization, however many fields were reorganized into different tables.

**Schema Design**

The star schema was explained in the excel workbook in detail. The following gives a high-level overview of the overall design.

**FACT\_SALES**: This main table maintains all the measures and includes relationships to customer, product, discounts. This design makes this table serves as a central index table to answer all our questions. For example., to answer the question ‘What engine is in Customer Smith’s car?’. Once we know Smith’s customer ID, we can directly query this table to find the VIN and from the VIN we can find the engine details. ORDERID serves as the primary key for this table.

To support analytics from different dimensions, I’ve created the following dimensions:

* **Customer**: This dimension tracks all the details about customer like first name, last name, address details, whether the customer is a repeat customer etc. In order to get clean analytics, creating a unique customer ID is important. Though in this exercise we are not de-duping the customer, creating a unique customerID key facilitates correlating sales with customer to track
* **Vehicle**: This dimension tracks details about a vehicle like make, model, year, color etc., VIN serves as the primary key for this table to uniquely identify each car.
* **Discount**: Discount is a dimension that will be very useful to understand which types of discount are selling cars with good margin etc., DiscountID is used the key to track the unique discount type like new customer etc.,.
* **Time**: Tracking the sales dimension gives the ability to inspect the sales transactions through the lens of time, to figure out how the sales have gone up/down over months/years. DATEID is created to uniquely identify the date dimension related attributes.

**Key Questions/Answers:**

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

Observing the problem statement and the supplied data indicates that the data is about Car Dealership. A relational schema is designed to solve the problem which gives us a STAR schema with Customer and Inventory as dimensions and Sales as Fact with the actual sale transactions that ‘refer to’ data from Customer and Inventory relations.

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

None of the data elements were left out except for some artificial columns. I’ve normalized the tables to remove duplicate information (customer’s address in two places etc.,), introduced two additional dimensions (time and discount) to facilitate analytics on the data from these perspectives. Some data is kept intentionally redundant (for e.g., REPEATCUSTOMER in CUSTOMER dimension so that customer service treat the repeated customers better).

* Why did you choose certain things as attributes? As keys?
  + I’ve ensured that every table has a unique primary key and the primary keys are linked to sales facts table. This facilitates the following:
* Removing duplicates
* Co-relate data between multiple data elements from different systems.
* Group things as a common key like customer first name and last name is identified with CUSTOMERID key
* From Customer, Sales, Inventory files many fields fit to be attributes naturally. Natural fitting data types were chosen for these attributes.
* What were the hardest decisions you had to make in this design process?
* Replicating “Repeat Customer” from Sales table to “Customer Relations” table.
* Creating time dimension. To solve this current problem, we do not need time dimension, but Time dimension is key for sales analytics.
* Creating discount types as a separate dimension.
* How does your schema design support data independence?
* The design is not depending upon the actual storage mechanism that will be used to store the data or the format in which it gets stored and retrieved.
* The design focusses on organizing data using abstract schema and data types.
* New data values can be easily added to individual tables. To support referential integrity new rows could be added to the FACT table.
* How may your schema design support the overarching goals of data curation (revisit objectives and activities of Week 1)?
* Organization – This schema organizes the data into different independent tables and facilitates connecting multiple tables through referential integrity.
* Storage – The suggested storage for storing this data was an RDBMS system that offers reliable data storage and recovery mechanisms.
* Discoverability – All the data is glued through the Sales table that can be used to easily co-relate relevant information.
* Access - Generally RDBMS system that hosts this data and schema supports multiple access and distributions patterns.
* Workflow: The workflow we’ve used to curate this data was described in detail in the Data Cleansing and Curation section
* Identification – The workflow process allows cross referencing data before importing the data into tables. (for e.g., a sale cannot be entered without a VIN etc.,)
* Reformatting: Some data was reformatted to different types (like normalizing discount types, creating Boolean for REPEATCUSTOMER etc.,)

* What are the pros and cons of your schema design?

Pros:

* + Normalized relational model.
  + Easy information retrieval -> Only few small queries required to answer any type of analytical question on the supplied data
  + Logical Independence and Physical Independence.

Cons:

* The current design fits well for the limited data that was presented. More comprehensive #of records use cases needed to improve the schema.
* Currently a customer is de-duped based on his address, Ideally, we may want to disassociate customer from address. To accomplish this, we need more data.
* Current schema assumes that every car should have a VIN#, if this is not the case (like international) current schema doesn’t supports this case.
* Which curation activities could enhance or sustain the database for future discovery and use for new purposes? What additional activities would you recommend?
* Strong data validation through Integration/ETL pipelines to ensure referential integrity
* Setting up data quality measures on each incoming file.
* Supporting UI, data discovery tools to observe and analyze the data.