CSI 4142 - Project Phase 1: Conceptual Design

Elli Traboulsi #300175764

Tamara Micic #300163312

Teodora Vukojevic #300199584

Group #4

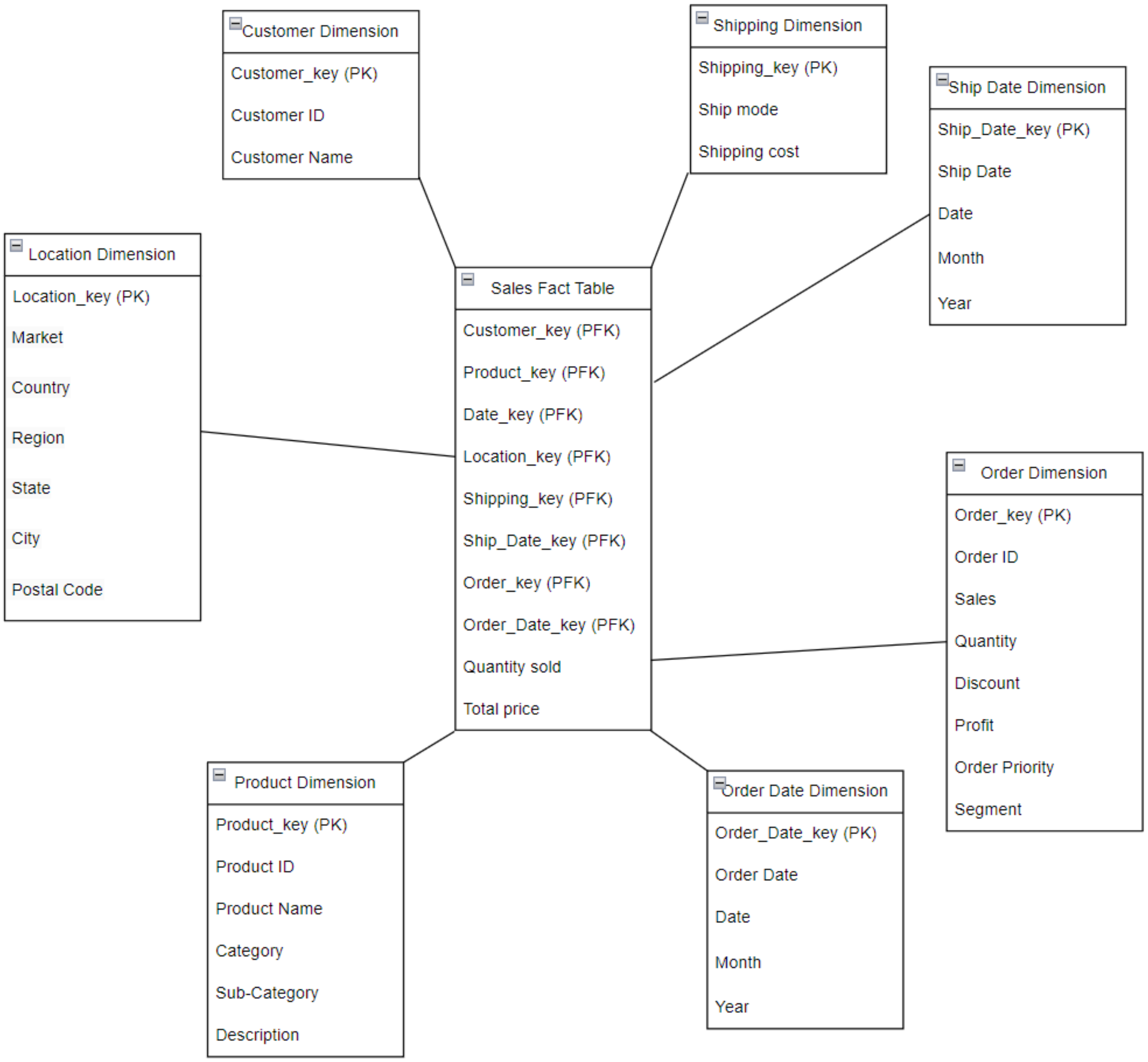
Prof. Yazan Otoum

TA Lansu Dai

9 February, 2024

Table of Contents

Conceptual Design of our Data Mart



Old dataset: https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci (Online Retail II UCI)
New dataset: https://www.kaggle.com/datasets/apoorvaappz/global-super-store-dataset (Global Super Store Dataset)

Questions to Address

1. What dimensions and attributes should be included to identify and track some trends in different terms over time?
The dimensions and attributes that should be included to identify and track some trends in different terms over time are location, date and product. For the dimension location, the

attributes used to track trends are postal code, market, city, region, state and country. Regarding the dimension date, the attributes used to track trends are time, day, month and year. For the dimension product, the attributes used to track trends are product name, category and sub-category.

2. Following 1, what key indicators do you need to store to obtain a clear picture of trends of those terms?
It is essential to store vital indicators in order to gain a comprehensive understanding of trends. These indicators include taking a look at the products that different countries purchase, examining purchases made at different times of the year (ex. holidays versus non-holidays), and looking at the products that individual customers purchase.

3. What are the external sources (if any) that can be potentially added to the dataset to enrich it further?
At first, we had only the Online Retail II UCI dataset, but replaced it with the Global Super Store dataset to enrich our conceptual design.

The Grain of Our Data Mart

Total price of an online order of a product(s) done on a day and shipped on a day to a customer in a certain location.

Dimensions and Dimensional Attributes

2. Detail all the dimensions and dimensional attributes. You should list the domains and show sample values. (e.g., Age: integer, minimum = 0 and maximum= 130, Sample value = 35).

Dimensions:
Customer Dimension

- Customer_key (defined by us)
    - Integer
    - Minimum = 00001
    - Maximum = 51291
    - Sample value = 40305
- Customer ID (Global SuperStore dataset):
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = JR-16210
- Customer Name (Global SuperStore dataset):

- String
- Minimum = N/A
- Maximum = N/A
- Sample value: Justin Ritter

Order Dimension

- Order_key (defined by us)
    - Integer
    - Minimum = 00001
    - Maximum = 51291
    - Sample value = 40455
- Order ID (Global SuperStore dataset)
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = ES-2013-1579342
- Sales (Global SuperStore dataset)
    - Float
    - Minimum = 0.444
    - Maximum = 22638.48
    - Sample value = 2735.952
- Quantity (Global SuperStore dataset)
    - Integer
    - Minimum = 1
    - Maximum = 14
    - Sample value = 10
- Discount (Global SuperStore dataset)
    - Float
    - Minimum = 0.0
    - Maximum = 0.85
    - Sample value = 0.2
- Profit (Global SuperStore dataset)
    - Float
    - Minimum = -6599.978
    - Maximum = 8399.976
    - Sample value = 358.02
- Order Priority (Global SuperStore dataset)
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = ‘Critical’
    - All values = 'Critical', 'Medium', 'High', 'Low'
- Segment (Global SuperStore dataset)
    - String

        - Minimum = N/A
        - Maximum = N/A
        - Sample value = 'Consumer'
        - All values = 'Consumer', 'Corporate', 'Home Office'

Order Date Dimension

- Order_Date_key (defined by us)
    - Integer
    - Minimum = 00001
    - Maximum = 51291
    - Sample value = 42502
- Order Date (Global SuperStore dataset)
    - DateTime
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = 29-10-2011
- Date (Global SuperStore dataset)
    - Integer
    - Minimum = 01
    - Maximum = 31
    - Sample value = 15
- Month (Global SuperStore dataset)
    - Integer
    - Minimum = 01
    - Maximum = 12
    - Sample value = 11
- Year (Global SuperStore dataset)
    - Integer
    - Minimum = 2011
    - Maximum = 2014
    - Sample value = 2013

Shipping Dimension

- Shipping_key (defined by us)
    - Integer
    - Minimum = 00001
    - Maximum = 51291
    - Sample value = 42501
- Shipping mode (Global SuperStore dataset)
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = 'Same Day'
    - All values: 'Same Day', 'Second Class', 'First Class', 'Standard Class'

- Shipping cost (Global SuperStore dataset)
    - Float
    - Minimum = 0.0
    - Maximum = 933.57
    - Sample value = 714.66

Ship Date Dimension

- Ship_Date_key (defined by us)
    - Integer
    - Minimum = 00001
    - Maximum = 51291
    - Sample value = 42502
- Ship Date (Global SuperStore dataset)
    - DateTime
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = 10-08-2012
- Date (Global SuperStore dataset)
    - Integer
    - Minimum = 01
    - Maximum = 31
    - Sample value = 16
- Month (Global SuperStore dataset)
    - Integer
    - Minimum = 01
    - Maximum = 12
    - Sample value = 09
- Year (Global SuperStore dataset)
    - Integer
    - Minimum = 2011
    - Maximum = 2014
    - Sample value = 2013

Product Dimension

- Product_key (defined by us)
    - Integer
    - Minimum = 00001
    - Maximum = 51291
    - Sample value = 32657
- Product ID (Global SuperStore dataset)
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = TEC-MA-10004125

- Product name (Global SuperStore dataset)
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = Brother Fax Machine, Laser
- Category (Global Super Store dataset)
    - Integer
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = Technology’
    - All values = 'Technology', 'Furniture', 'Office Supplies'
- Sub-category (Global Super Store dataset)
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = Bookcases

Location Dimension

- Location_key (defined by us)
    - Integer
    - Minimum = 00001
    - Maximum = 51291
    - Sample value = 11278
- Market (Global Super Store dataset):
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = ‘US’
    - All values: 'US', 'APAC', 'EU', 'Africa', 'EMEA', 'LATAM', 'Canada' (all values, meaning United States, Asia-Pacific, European Union, Africa, Europe/Middle East/Africa, Latin America, respectively)
- Country (Global SuperStore dataset):
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = “United Kingdom”
    - There are 147 unique countries
- Region (Global SuperStore dataset):
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = ‘East’
    - All values: 'East', 'Oceania', 'Central', 'Africa', 'West', 'South', 'Central Asia', 'EMEA', 'North Asia', 'North', 'Caribbean', 'Southeast Asia', 'Canada'

- State (Global Super Store dataset):
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = 'New York'
    - There are 1094 unique states
- City (Global Super Store dataset):
    - String
    - Minimum = N/A
    - Maximum = N/A
    - Sample value = ‘Brisbane’
    - There are 3636 unique cities
- Postal Code (Global Super Store dataset):
    - Integer
    - Minimum = 1040
    - Maximum = 99301
    - Sample value = 55901
    - There are 632 unique values

Measure/Fact Details

3. Detail all the measures/facts. You should list the domains and sample values. (e.g., Age: integer, minimum = 0 and maximum = 130, Sample value = 35).

- Quantity sold (equivalent to ‘Quantity’ column):
    - Integer
    - Minimum = 1
    - Maximum = 14
    - Sample value = 12
- Total price (equivalent to ‘Sales’ column):
    - Integer
    - Minimum = 0.444
    - Maximum = 22638.48
    - Sample value = 4630.4755

4. Remember to detail all your assumptions.

- We assumed that the currency is USD for all monetary values
- We assumed what the Market values mean (ex. We assumed LATAM means Latin America)

Checklist of “10 design mistakes”

5. Make a checklist (Use Tables) of the “10 design mistakes” mentioned at the end of Module 1, Part C and show how you avoided/handled those mistakes (Where Applicable).

	Design Mistake	How we handled it/Comments	Checklist
1	Placing text attributes in the fact table	We made sure all attributes in the fact table were numerical attributes.	☑
2	Limiting verbose descriptions to save space	N/A	☑
3	Normalizing to save space (leads to slower queries)	N/A	☑
4	Ignoring the need to track changes	We did not make any changes therefore no tracking was made. All assumptions were detailed.	☑
5	Adding new hardware to solve all query performance issues	N/A	☑
6	Using operational keys as the primary keys	We create our own new keys for each dimension instead of using IDs in the datasets (ex. We create Customer_key, even though we also have Customer ID).	☑
7	Neglecting to declare (and comply with) the grain	We made sure to clearly declare the grain.	☑
8	Neglecting a detailed design	We ensured our design was as clear as possible. We did not spare any details.	☑
9	Expecting users to query normalized data (3 again)	N/A	☑
10	Failing to conform facts and dimensions	We made sure that we did not have mismatched, inconsistent, or incomplete data between facts and dimensions	☑

Work Summary

6. A summary of your team's work plan, including the times and dates you met, how you divided the work, and how you often meet with the TA.

Meeting Summaries

Team Meeting 1
Date: February 1st, 7pm-9pm

Topics of Discussion:

- Declared the grain of our data mart
- Creating a first draft of the individual sales transaction schema
    - Discussions about what should be included in the fact table, what the dimensions and attributes should be.
- Using the "10 design mistakes" as a guideline
- Decided to use draw.io for our conceptual schema

Questions for the TA:

- How much detail?
    - For the date key for example. Can we derive some data, for example, day of the week, continent, total price.
    - Answer: we can do this.
- Should we create a dimension for the transaction or should all of this information be in the grain?
    - Answer: no dimension
- All of the columns in our dataset have duplicate values, is that ok?
    - Should we create a "Transaction number" column with a unique number (1-500,00+) for all rows?
    - Don't need to add since it doesn't bring any value to the data mart
- Can attributes in the fact table be left out of the grain? Yes
- Fact table name? Transaction or Invoice or Sales
- Geography dimension with one attribute, Country?

Meeting with TA Lansu
Date: February 4th, 10am-10:30am

Meeting Summary:

- Answers to the above questions.
- "Online sales from a country" for the grain
- Find at least one other dataset and add dimensions from those datasets into the schema

Team Meeting 2

Date: February 6th, 3pm-5pm

Topics of Discussion:

- Adding the second dataset to the conceptual design
- Dimension attribute descriptions
- Measure/facts attribute descriptions

Team Meeting 3

Date: February 9th, 10am-10:30am (after class with the prof)

We talked to the professor after class and realized our second dataset was better than our original one. And since it has a greater amount of features, we should use the second dataset only. The first dataset would not enrich our data mart so we are not keeping it. He said having other documents is not necessary.

- Updating the conceptual schema

Work Distribution

As a group we:

- Defined the grain
- Discussed and created a rough draft of our conceptual schema
- Looked for a second dataset and opted to use this new one instead of our original one

Tamara:

- Created and adjusted the schema
- Detailed dimensional attributes
- Detailed measures/facts

Elli:

- Documented the work summary and communicated with the TA through email
- Detailed dimensional attributes
- Went through the “10 design mistakes” checklist

Teodora:

- Answered questions 1 and 2 in the “Questions to Address”

References

Du Mortier, G. (2023, August 29). *How to create a data model for a Data Warehouse*. Vertabelo Data Modeler. https://vertabelo.com/blog/warehouse-data-model/

How to enter the APAC market: Tips for Asia-Pacific localization. Smartling. (n.d.). https://www.smartling.com/resources/101/apac-localization/#:~:text=What%20is%20the%20APAC%20region,%2C%20Southeast%20Asia%2C%20and%20Oceania.

LATAM definition and meaning | Collins english dictionary. (n.d.). https://www.collinsdictionary.com/dictionary/english/latam