CIS 355 midterm review

Fall 2014
Dr. Joseph Clark

Overview of exam

- 20-25 multiple choice questions on material in slides and readings
- 2 dimensional modeling exercises

Topic areas to review

- Databases—what, and why?
- Differences between transactional/operational and informational/analytical database systems.
- Dimensional model development process
- Data warehouse design considerations
- Kimball architecture and agility

Databases—what and why?

- What are the problems with just using files?
- Program-data independence
- Conceptual modeling: how to read an ER diagram
- How data storage works, and what physical constraints are relevant.
- What's a database index?
- Slide deck intro_to_databases
- Slide-deck physical_database_design

Transactional vs Analytical systems

- Trade-off between “SELECT” and “INSERT/UPDATE/DELETE” performance. (aka “read” and “write” performance)
  - This is why you need to know about indexes.
- Different users, design goals, and characteristics of the two database types.
- Slide deck physical_database_design
- Slide deck intro_data_warehousing
- Kimball chapter 1

Dimensional Model Development Processes

- Identifying and understanding business processes (ch.3) and value chains (ch.4)
- Steps in dimensional modeling (ch.3)
- Organizational considerations (Guthy-Renker case)
- Different types of questions that can be answered (ch. 4 slides)
- Kimball chapter 3
- Kimball chapter 4
- Guthy-Renker case
- Slide decks retailsales_intro & inventory_ch4

DW/BI design considerations

- Facts and dimensions; grain
- Additive, semi-additive, non-additive facts
- Three types of fact tables
- Difference from 3NF normalized models (i.e. why dimension tables have so many darn attributes)
- Kimball ch. 3
- Kimball ch. 4
- Slide decks retailsales_intro & inventory_ch4

Kimball architecture & agility

- Conformed dimensions
- Enterprise bus architecture (and bus matrix)
- Key idea: re-use of previously-done work for (1) consistency and (2) faster work
- What “Agile” and “iterative” mean in software development
- How Kimball's architecture contrasts with Inmon's corporate information factory architecture and why
- Kimball ch. 1 & ch. 4
- Slide deck inventory_ch4

Note on Chapter 2

- Kimball's chapter 2 is a sort of preview of all topics in the book.
- It might be helpful for review.

Dimensional modeling exercises

- Exercises will give you a case study, for example:

A university bookstore sells a lot of apparel—t-shirts, hats, sweaters, etc—and would like to better understand purchasing patterns. They suspect that sales may be affected by the rhythms of university life, for example, different types of products might be bought at the beginning of the semester and the end, before and after football games, etc. They have multiple “storefronts” including the main campus bookstore and an online catalog. Products may be sliced a number of ways including by manufacturer, by size, material, and so on.

- Read it and carry out four tasks... (next slide)

Dimensional modeling exercises

1) Draw a dimensional model to meet this organization's analytical needs as described in the case.

   Include attributes in the dimension tables that would be important to this case and use "..." for other attributes that may be obvious or universal.

2) Declare the grain of the fact table:

3) Describe three queries that would likely provide valuable information to the organization.

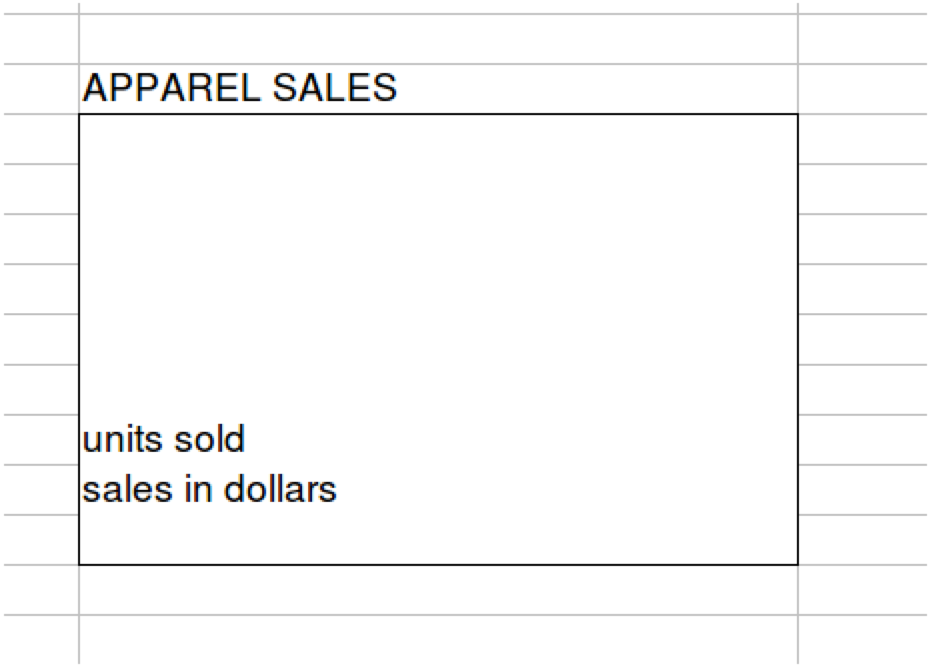
   In English or SQL, as you prefer.

4) Sketch out an example "report" that one of the queries would provide.

Demonstration

- I tried to do this with a video screen capture, but it crashed several times, so just follow me step by step on these slides.
- I'm using a spreadsheet here, but you can do this with pen and pencil.
- (next slide)

Start with a fact



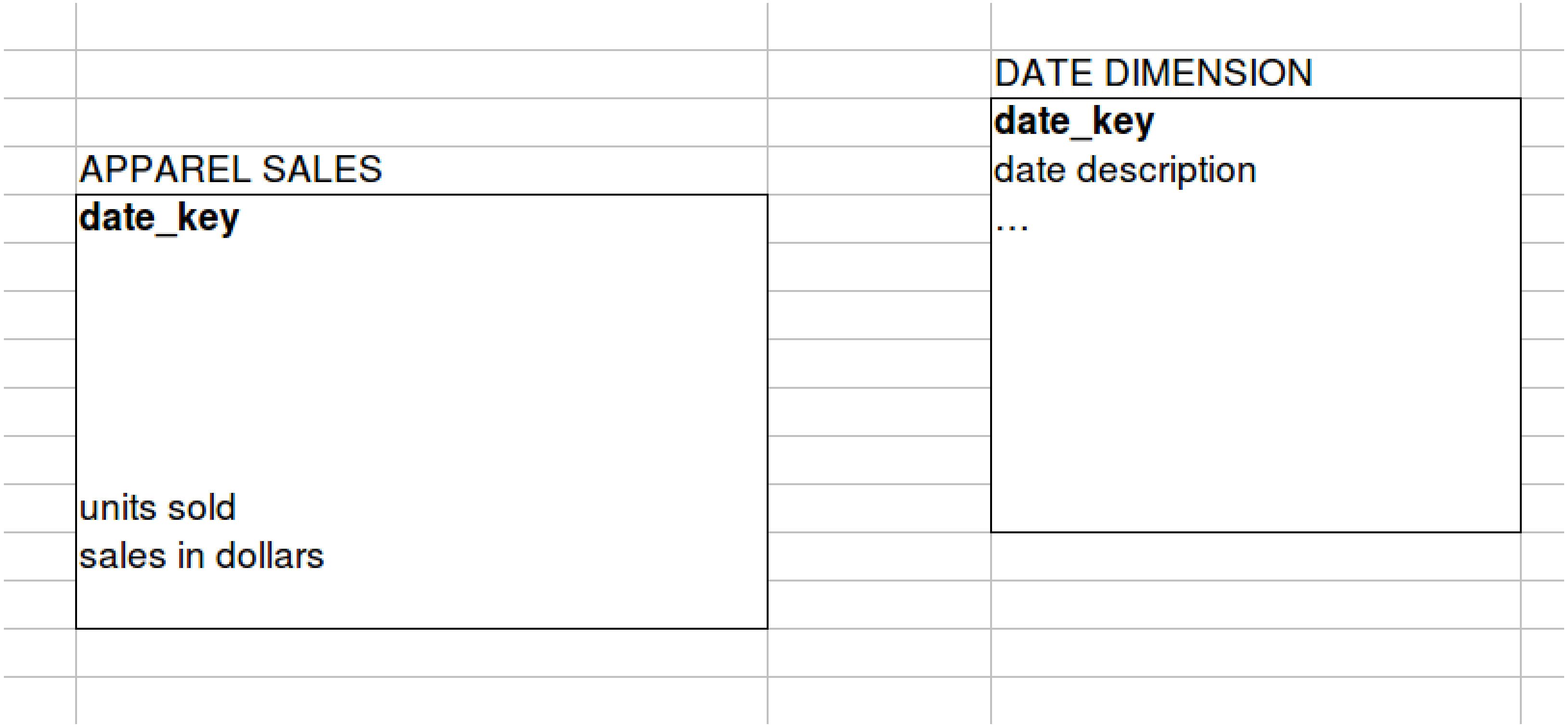
APPAREL SALES

units sold

sales in dollars

Add a date dimension

- Pretty much every dimensional model has a date dimension.
- You can use "..." to stand for all the obvious or "usual" things such as are found in chapter 3's date dimension.
- Don't forget to create the key relationship. (On paper, you'd draw an arrow connecting the FK in the fact table to the PK in the date dimension.)



APPAREL SALES

date_key

units sold

sales in dollars

DATE DIMENSION

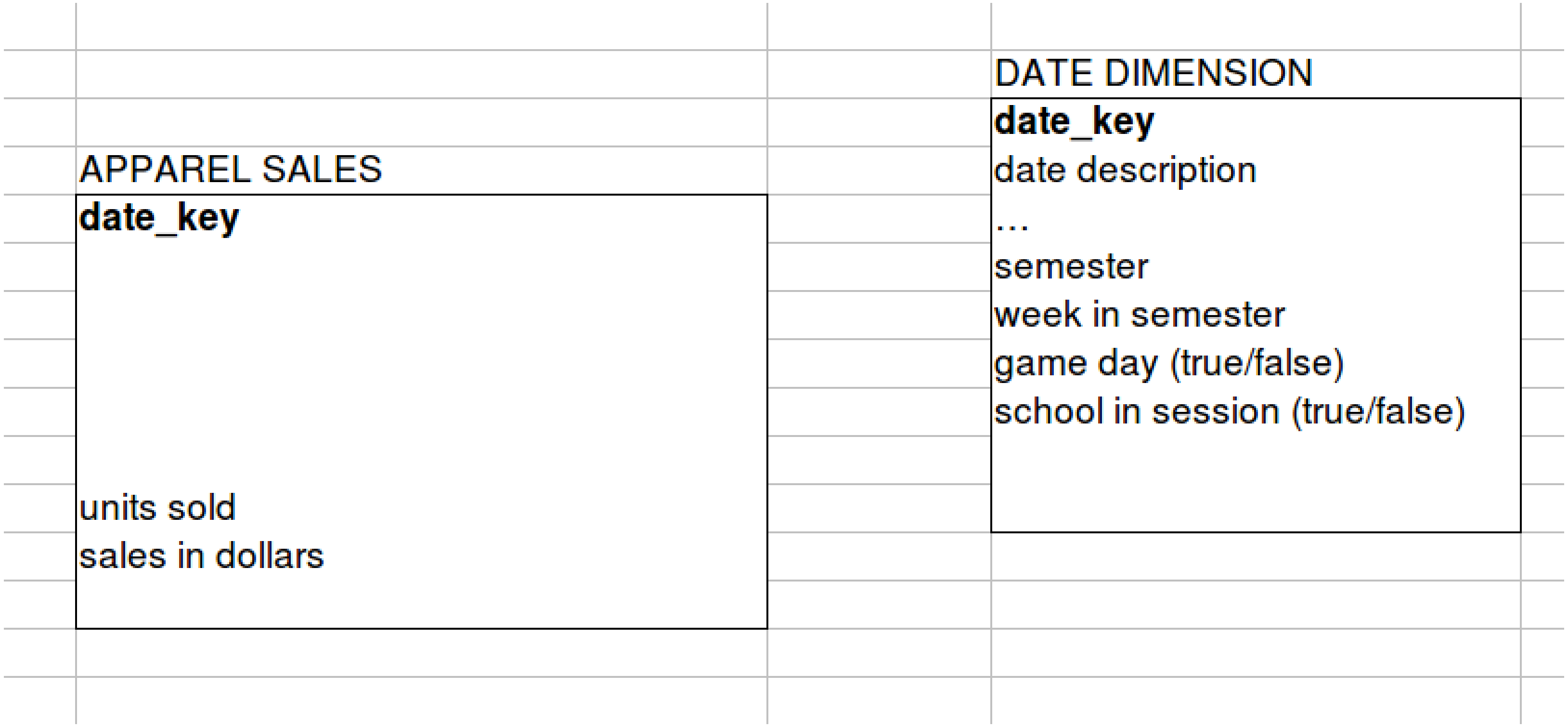
date_key

date description

...

Now enrich the dimension

- In addition to the “usual stuff”, there are some unique things in this case. They want to analyze data by the rhythms of the academic calendar, so let's add a few more attributes of dates.



APPAREL SALES

date_key

units sold

sales in dollars

DATE DIMENSION

date_key

date description

…

semester

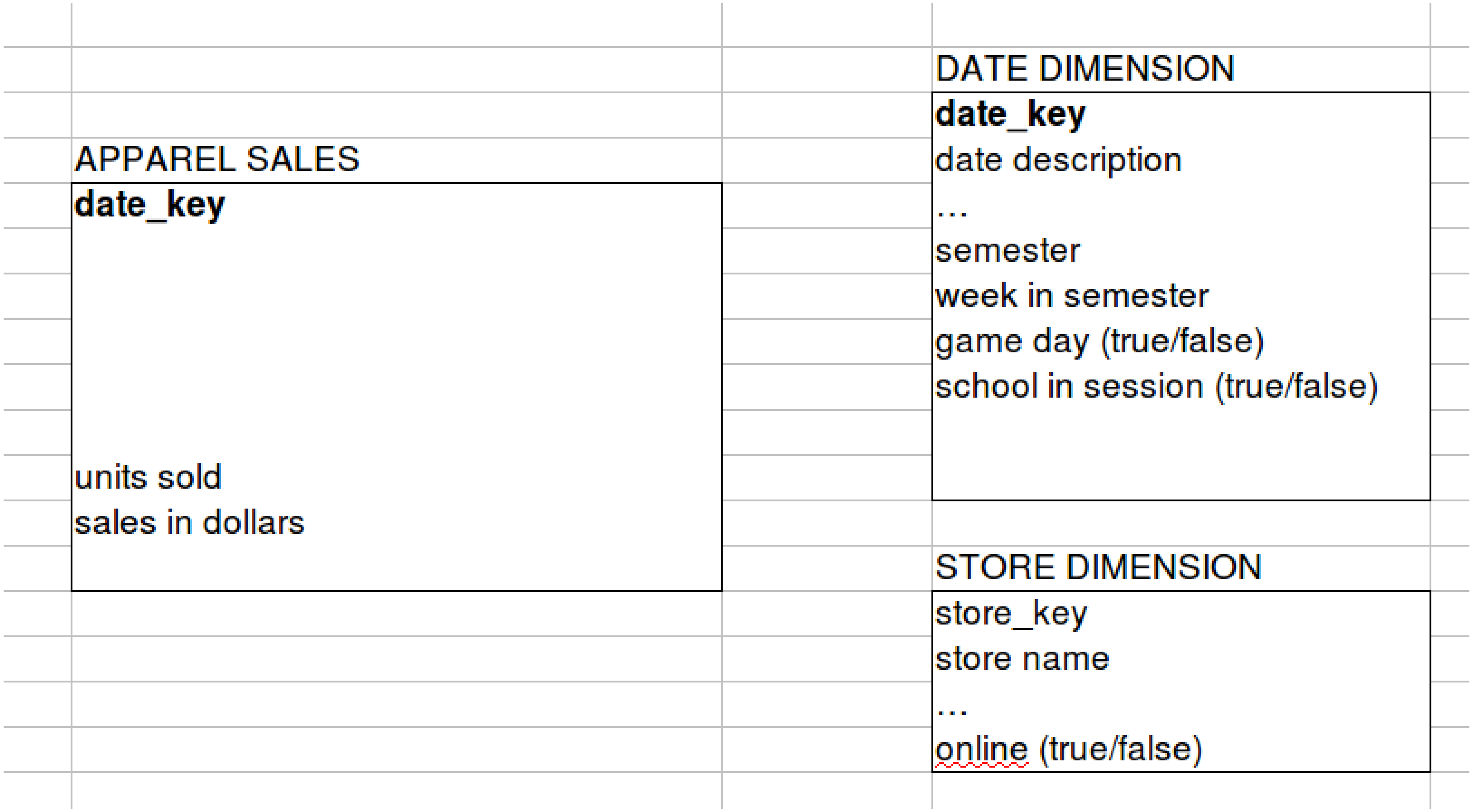
week in semester

game day (true/false)

school in session (true/false)

Store dimension

- This one is a no-brainer. One thing specifically mentioned in the case is online vs offline storefronts.



APPAREL SALES

date_key

units sold

sales in dollars

DATE DIMENSION

date_key

date description

…

semester

week in semester

game day (true/false)

school in session (true/false)

STORE DIMENSION

store_key

store name

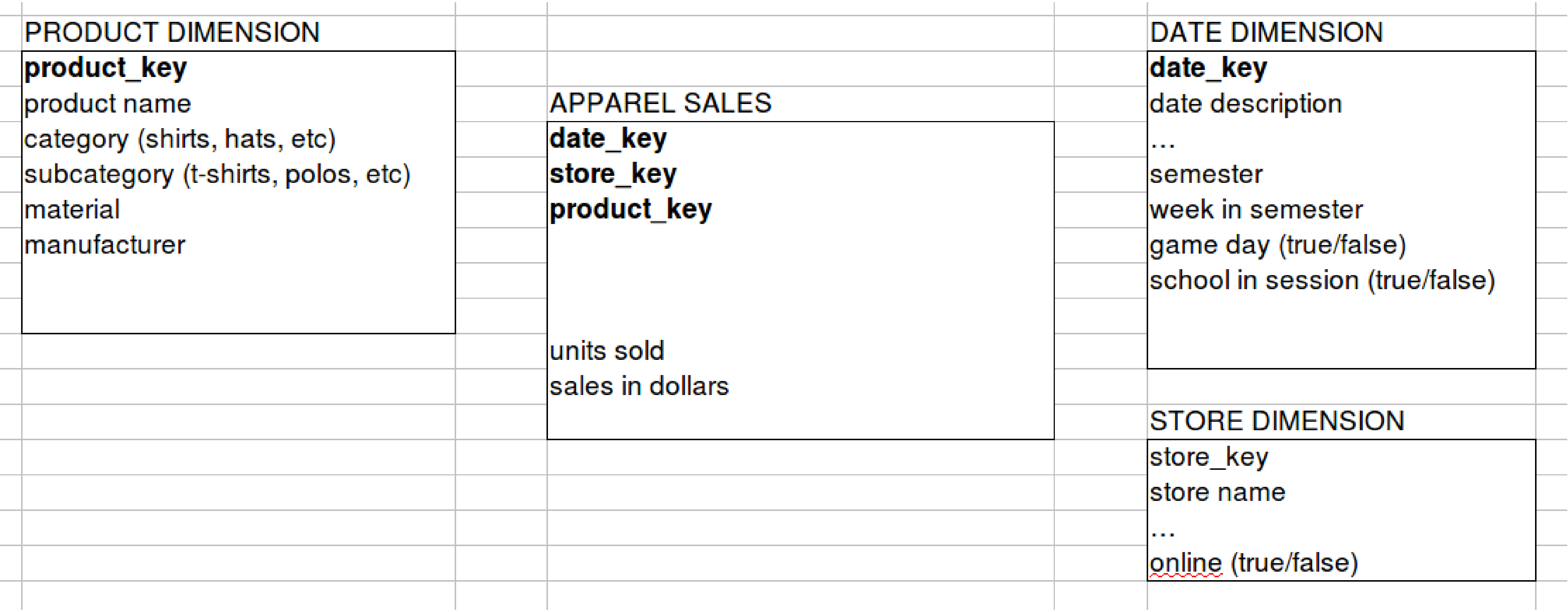
…

online (true/false)

Product Dimension

- This is a model specifically for apparel, and the case gives you a few things to include.
- You can use parentheses to give me a sample of what would go into an attribute, if you don't think it's clear from the name. In the example (next slide) I give examples of what would be a “category” and a “subcategory”.
- Would manufacturer be an attribute of Product, or a dimension of its own? It's a judgment call. In this case, I'm guided by the principle that we don't want to explode this model into too many tables. That would increase complexity, making it harder for business users to use this model, without improving performance.
- (next slide)

Product dimension added



PRODUCT DIMENSION

product_key
product name
category (shirts, hats, etc)
subcategory (t-shirts, polos, etc)
material
manufacturer

APPAREL SALES

date_key
store_key
product_key

units sold
sales in dollars

DATE DIMENSION

date_key
date description
…
semester
week in semester
game day (true/false)
school in session (true/false)

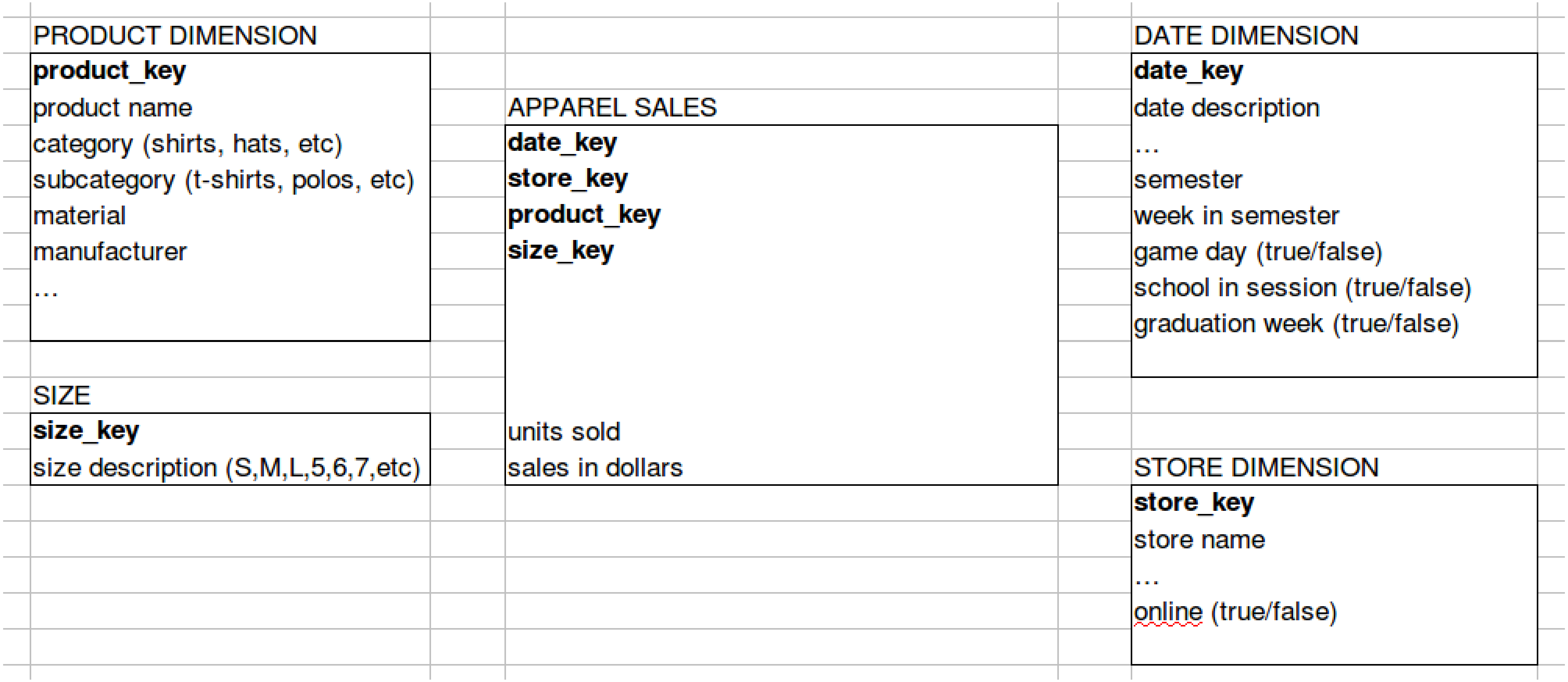
STORE DIMENSION

store_key
store name
…
online (true/false)

What about size?

- Why not include “size” in the product dimension?
- Think about this: there may be dozens of sizes for some items (like mens shirts with various neck and sleeve lengths). If there are a few hundred items in the store, this could explode the dimension to thousands of rows.
- Instead, let's create a size dimension.
- (next slide)

With a size dimension



PRODUCT DIMENSION

product_key
product name
category (shirts, hats, etc)
subcategory (t-shirts, polos, etc)
material
manufacturer
…

SIZE

size_key
size description (S,M,L,5,6,7,etc)

APPAREL SALES

date_key
store_key
product_key
size_key

units sold
sales in dollars

DATE DIMENSION

date_key
date description
…
semester
week in semester
game day (true/false)
school in session (true/false)
graduation week (true/false)

STORE DIMENSION

store_key
store name
…
online (true/false)

Size is a mini-dimension

- We created a separate size dimension to keep the product dimension from exploding.
- Another case where you often see this is a “Time of Day” dimension. If you were to include this with “date” in a Date-Time dimension, you could end up with half a million rows a year (one per minute). If you separate these, you'll instead have:
  - 1440 rows in your Time dimension, one per minute
  - 365 rows in your Date dimension per year
- So you save a lot of space and headache by doing this.

Step 2: state the grain

- Grain can be stated “one row per...”
- Our grain here is one row per sale of a product. (Not one row per transaction, because a transaction can include more than one product.)

Step 3: suggest some queries

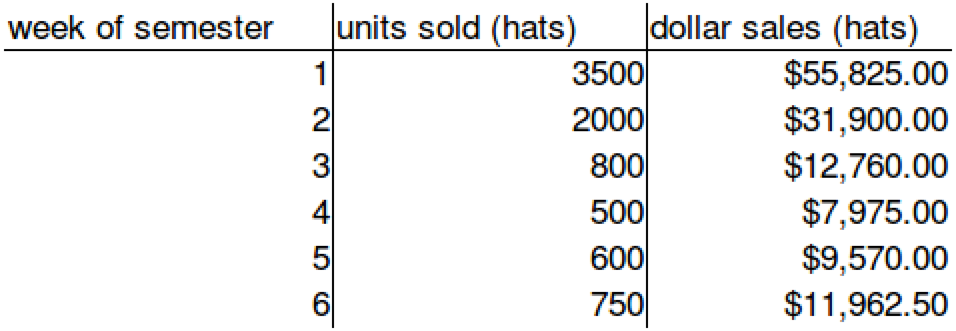
- The purpose of this task is to make sure you know what a dimensional model is for. They need not be complex queries. They will likely include the word “by”.
- For example, we could query here:
  - Sales of hats by week of the semester.
  - Sales in 2014 on game days vs non-game days.
  - Sales of t-shirts by size, by semester.

Step 4: mock report

- This is again to show that you understand how facts and dimensions translate into analyses that you could run.
- Typically, dimensions give you most of the headers of the report. Adding more dimensions means you get a more detailed report. The fact(s) are usually the rightmost column(s) of the report.
- (next slide)

Step 4: mock report

- Just a few rows of made-up data will do the job.



week of semester	units sold (hats)	dollar sales (hats)
1	3500	$55,825.00
2	2000	$31,900.00
3	800	$12,760.00
4	500	$7,975.00
5	600	$9,570.00
6	750	$11,962.50

Questions?

- If there are any questions these slides and my video haven't answered, feel free to e-mail:
  - joseph.w.clark@asu.edu