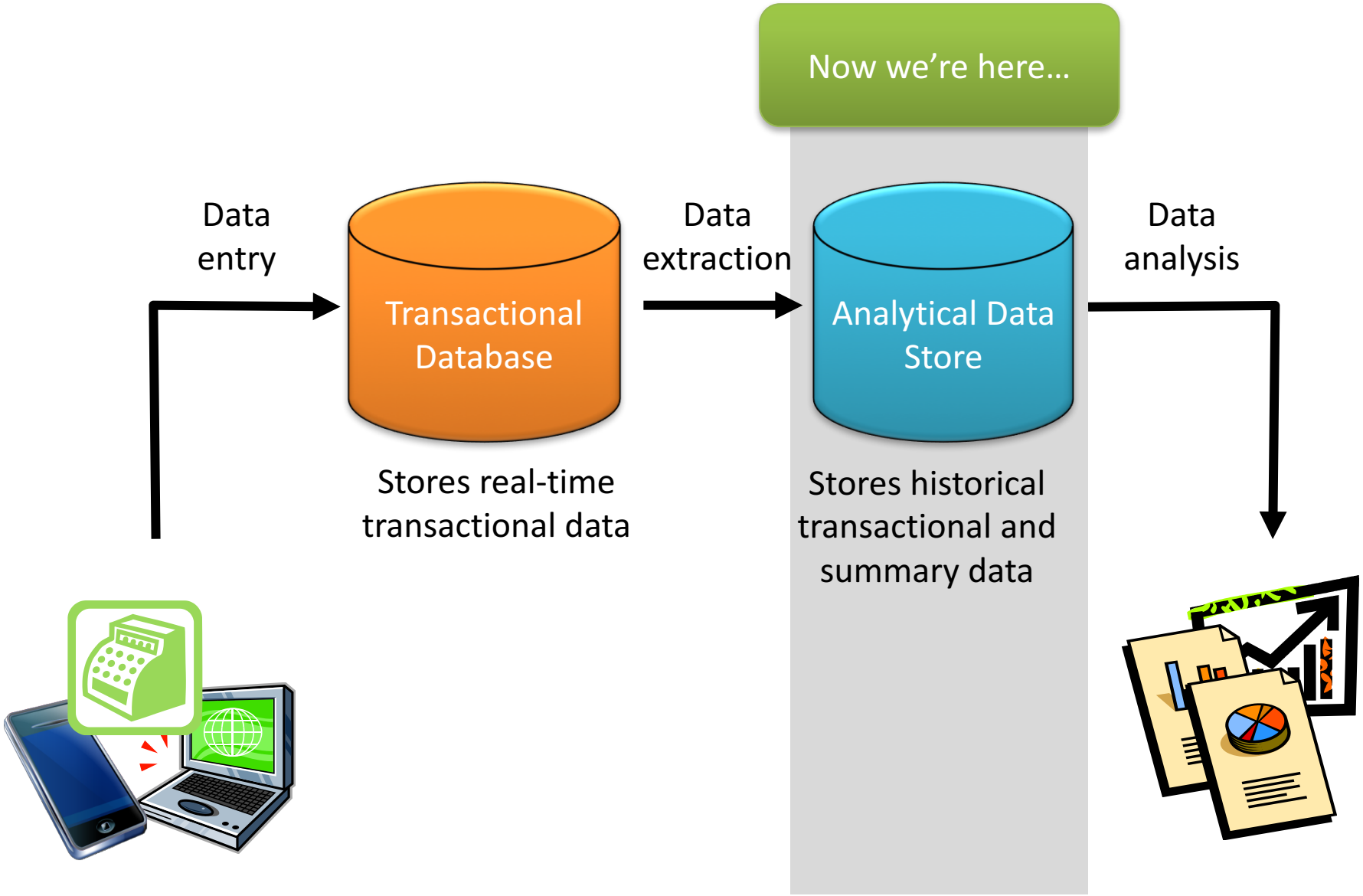
MIS2502:
Data Analytics
Dimensional Data Modeling

Alvin Zuyin Zheng
zheng@temple.edu
http://community.mis.temple.edu/zuyinzheng/

Where we are…


Now we're here…
Data entry
Transactional Database
Stores real-time transactional data
Data extraction
Analytical Data Store
Stores historical transactional and summary data
Data analysis

What do we know so far?

Why are relational databases good for storing transaction data?

Why are they bad for analytical processing?

What’s the solution?

Some terminology

Data Warehouse

- Takes many forms
- Really is just a repository for historical data

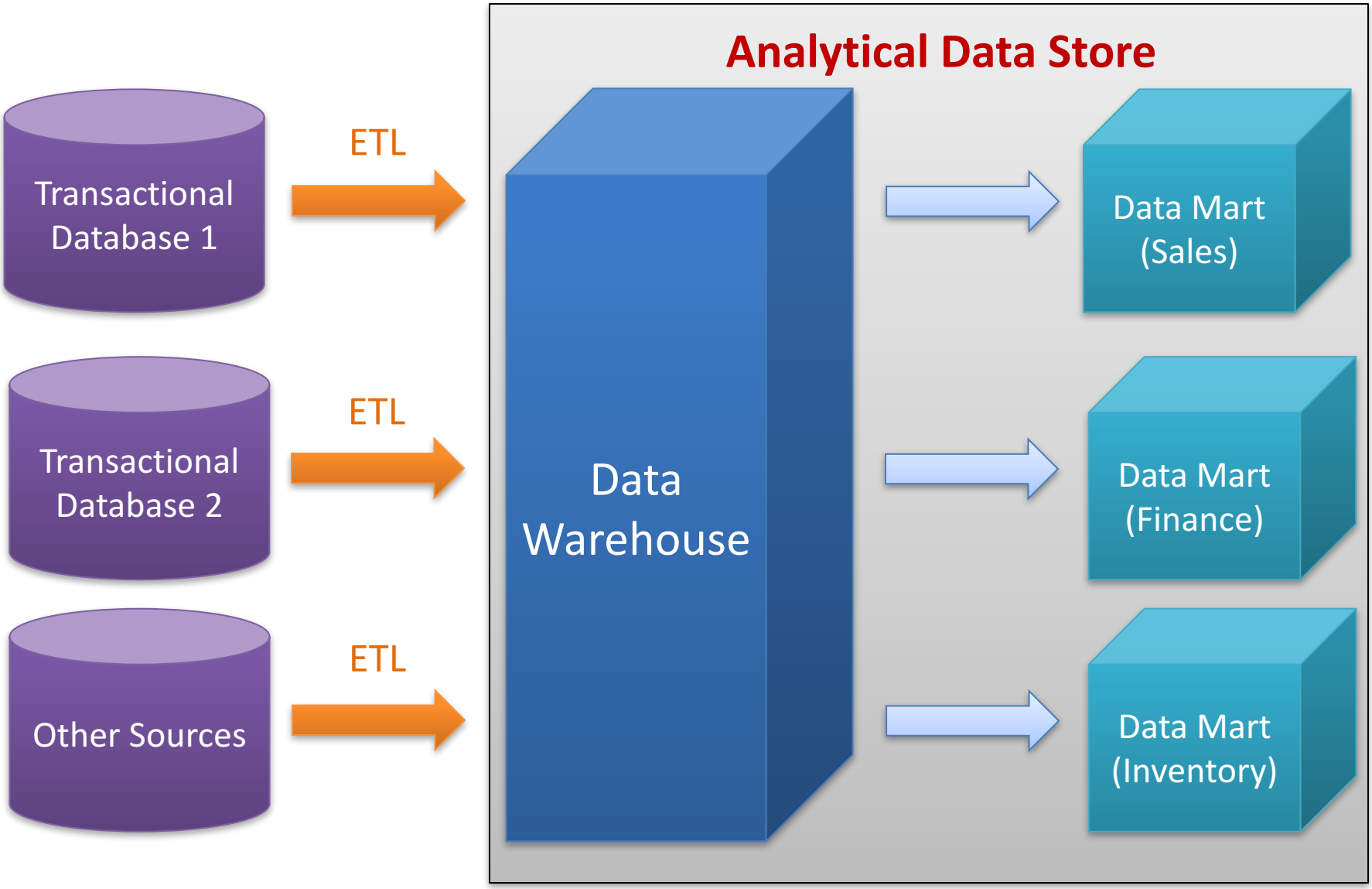
Data Mart

- **Subset** of the Data Warehouse
- Designed for specific analysis

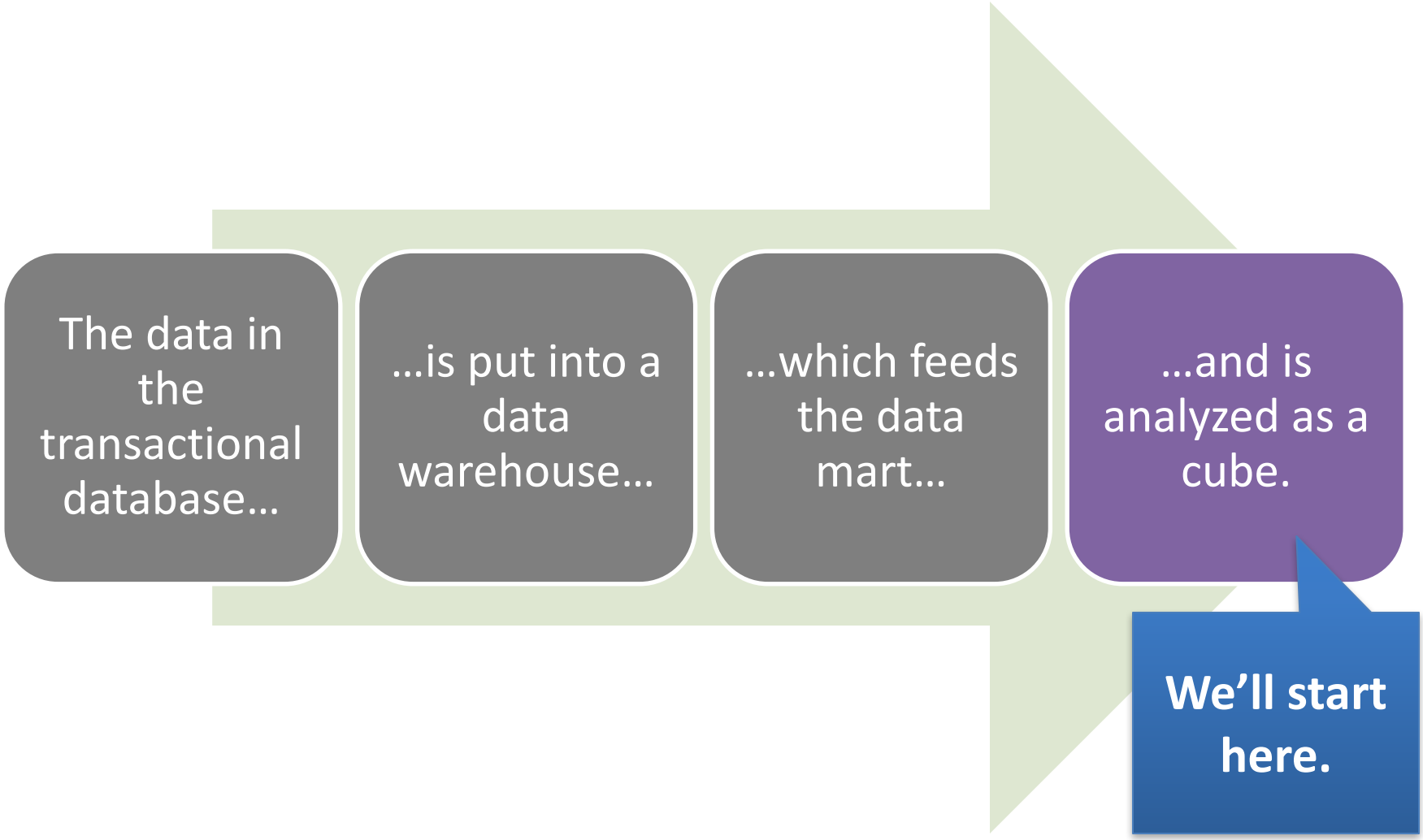
Data Cube

- Organization of data as a "multidimensional matrix"
- Implementation of a Data Mart

The Actual Process

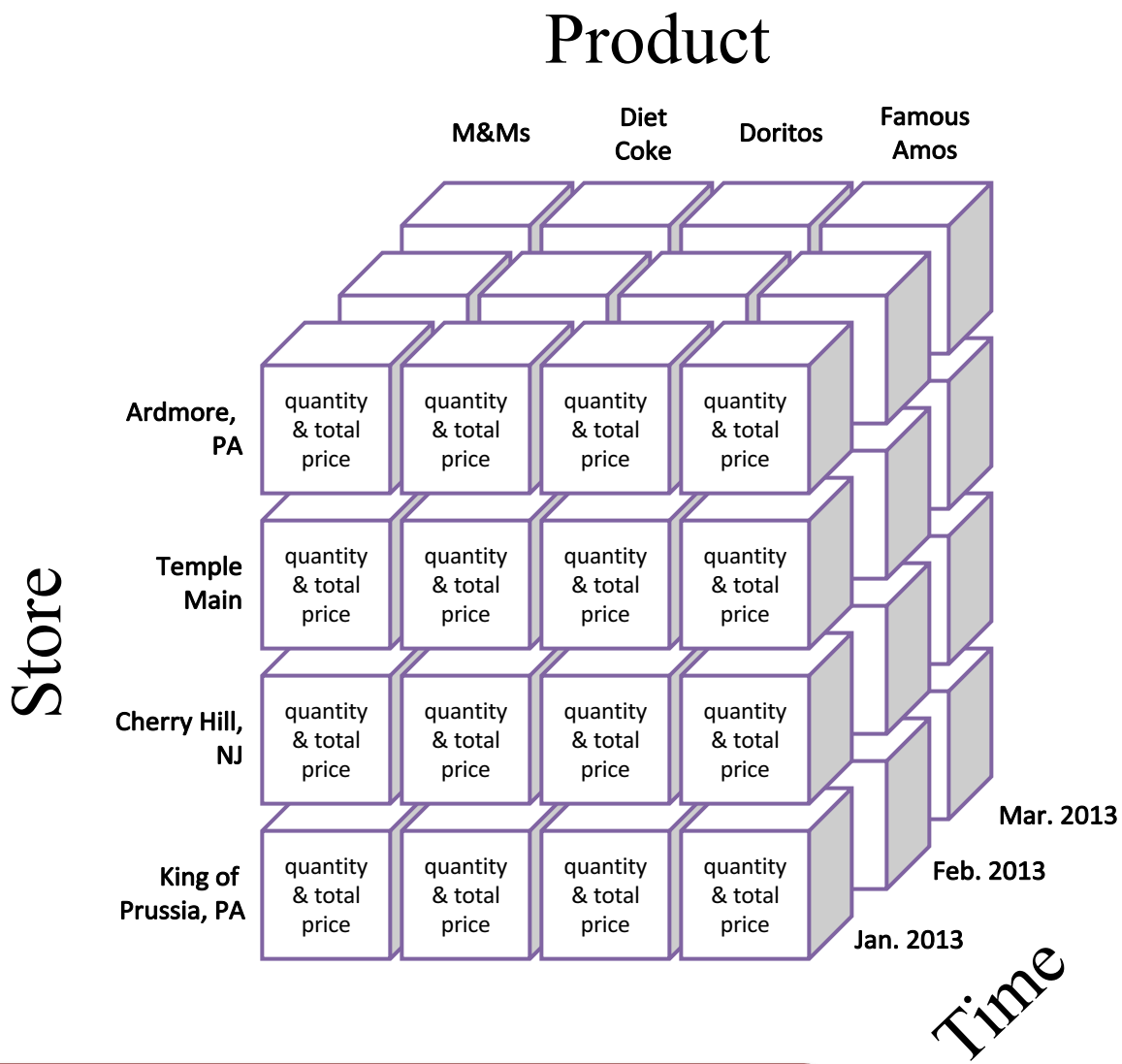


How they all relate



The Data Cube

- Core component of Online Analytical Processing (OLAP) and Multidimensional Data Analysis
- Made up of "facts" and "dimensions"

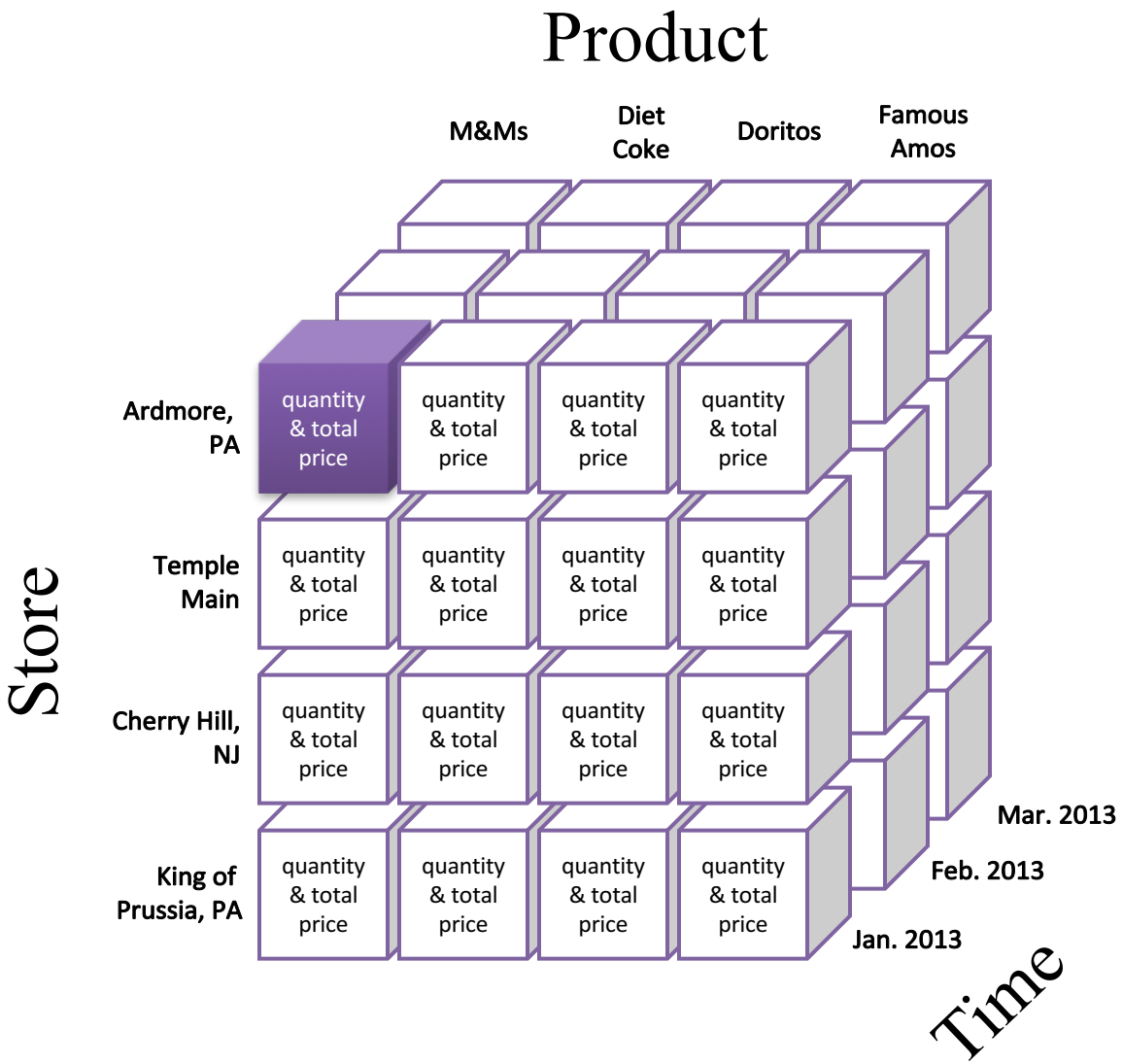


Quantity sold and total price are measured facts.
Why isn't product price a measured fact?

The Data Cube

The highlighted element represents all the M&Ms sold in Ardmore, PA in January, 2013

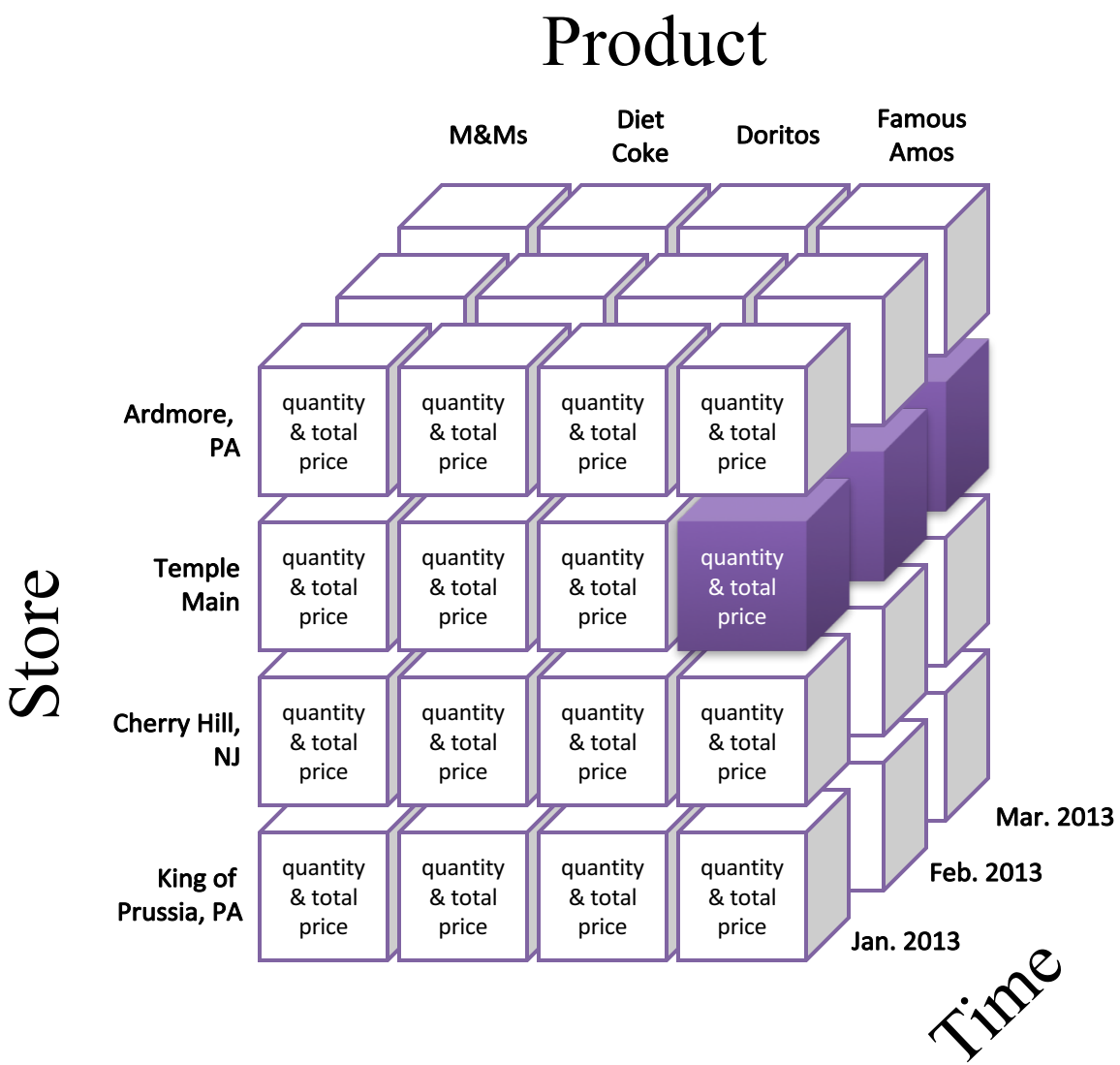
A single summary record representing a business event (monthly sales).



Slicing the Data

The highlighted elements represent Famous Amos cookies sold on Temple's Main campus from January to March, 2013

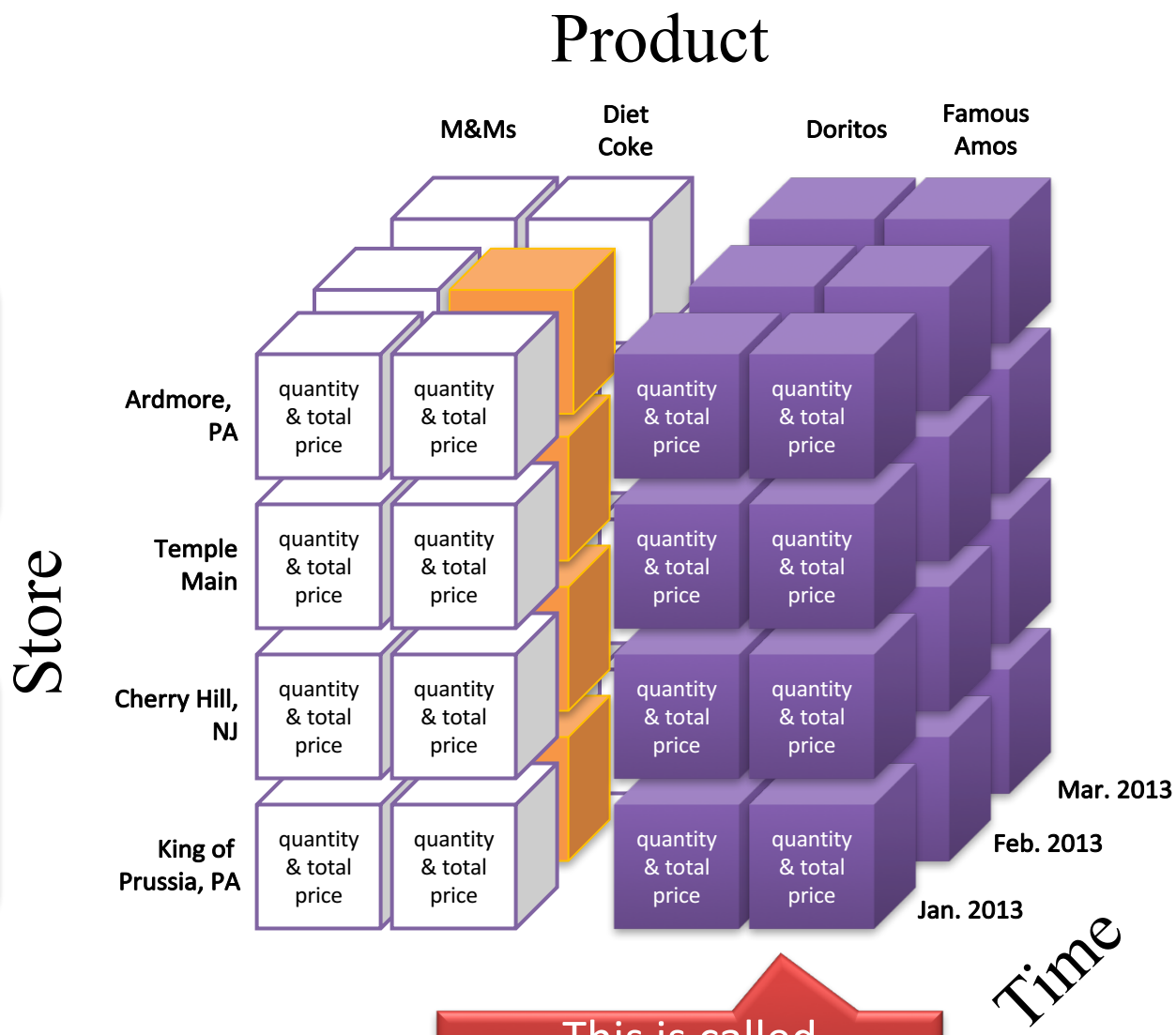
This is called "**SLICING** the data."



Dicing the Data

What do the orange highlighted elements represent?

What do the purple highlighted elements represent?



This is called "**DICING** the data"

Could you have a data mart with five dimensions?

Then why does our example (and most others) only have three?

Modeling a data cube: The Star Schema

Transactional databases aren't built around dimensions

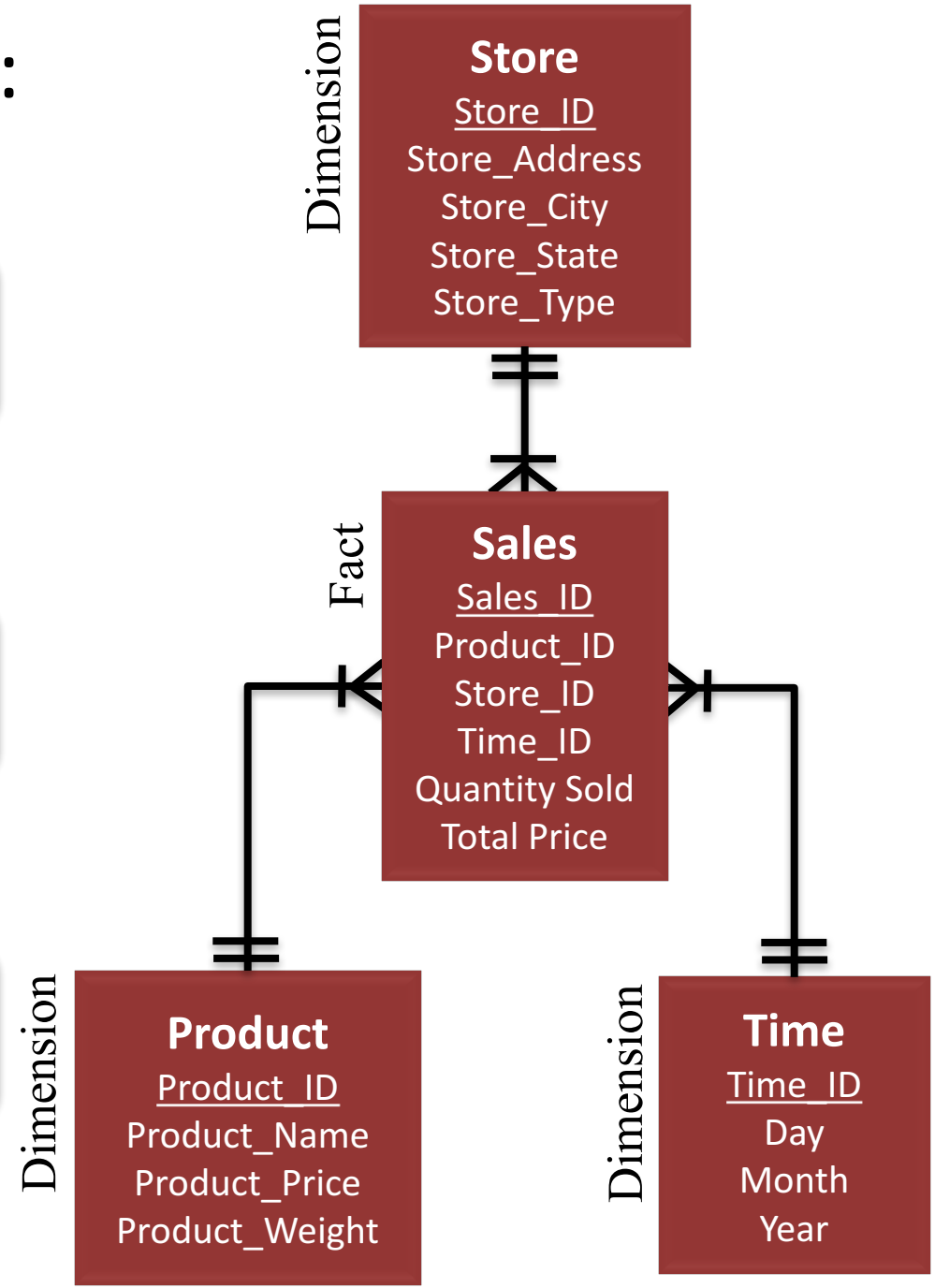
- They don't map well to cubes
- They aren't set up for summarization

So we build a star schema

- Built around "dimensions" and "facts"
- Simplified relational model

The star schema facilitates

- Aggregating individual transactions
- Creation of cubes



Fact Table

- Contain facts (numeric measurements) associated with a specific business process
- Contain foreign keys that refer to dimension tables

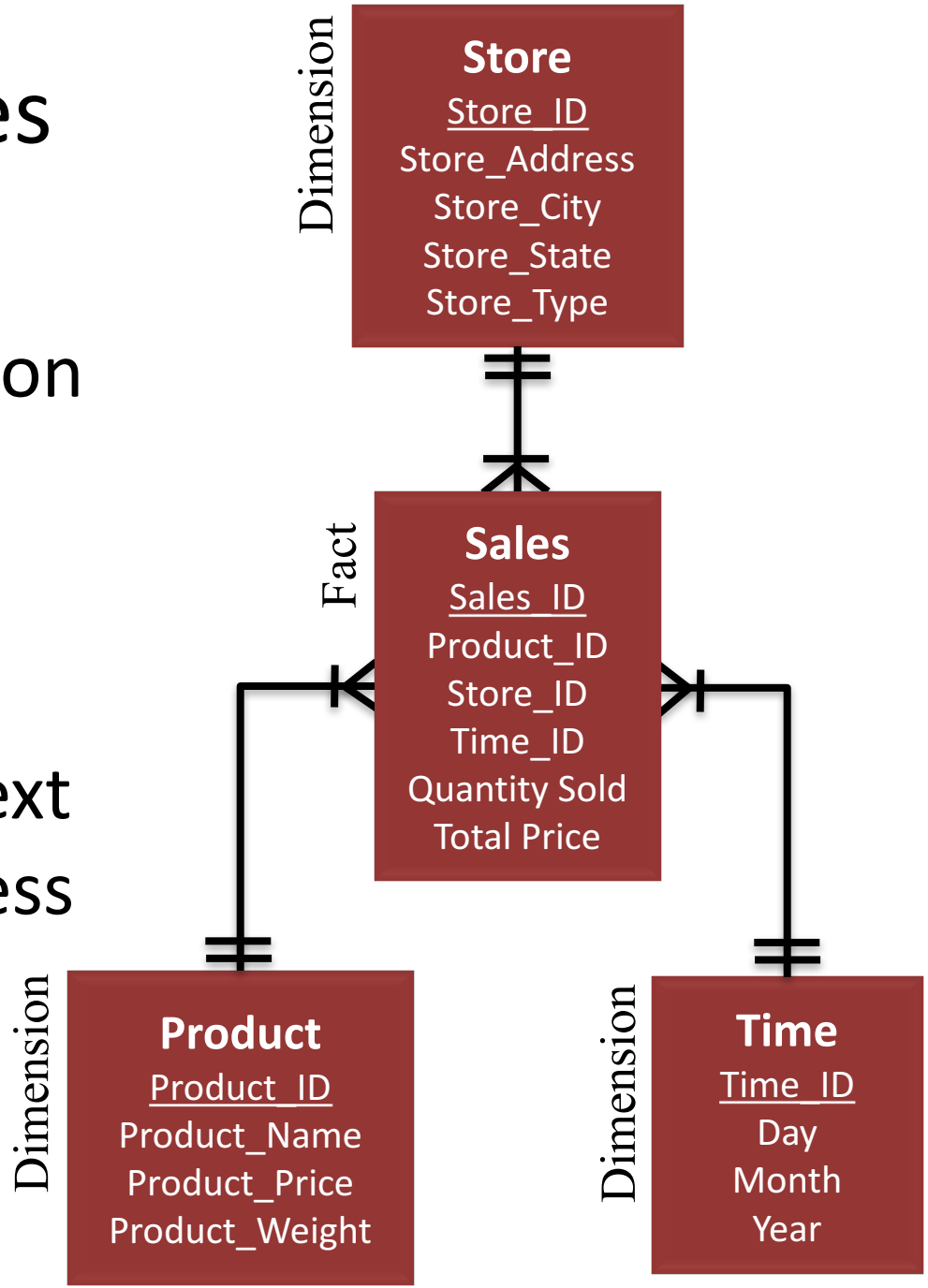
Fact

Sales

Sales_ID
Product_ID
Store_ID
Time_ID
Quantity Sold
Total Price

Dimension Tables

- Contain text and descriptive information
- Provide the "who, what, where, when, why, and how" context surrounding a business process event

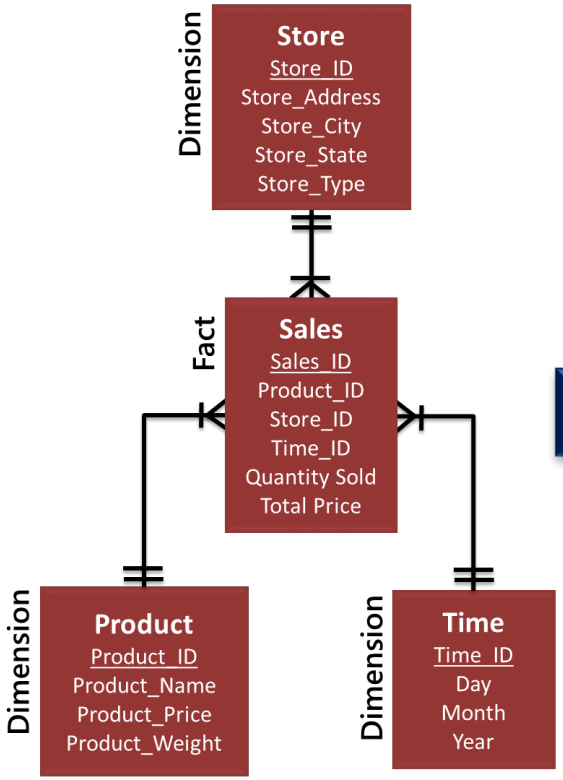


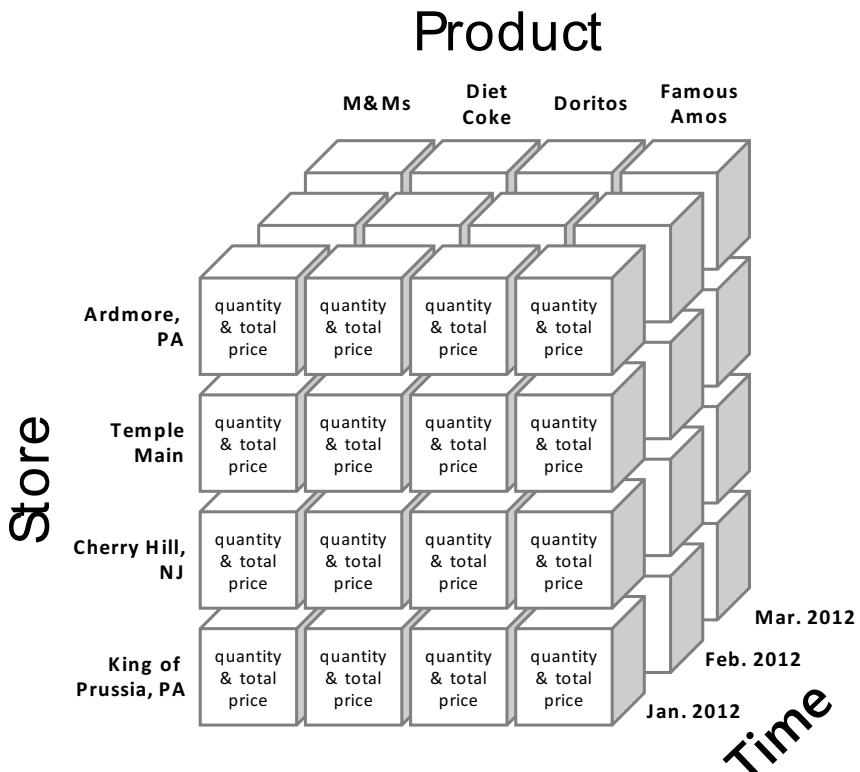
From Star Schema to Data Cube

A Cube typically uses a Star Schema as its source

and stores precomputed summarized (aggregated) data

Much more efficient, but can't be changed (non-volatile)





Designing the Star Schema

Kimball's Four Step Process for Data Cube Design
(Kimball et al., 2008)

1. Choose the business process

2. Identify the fact

3. Decide on the level of granularity

4. Identify the dimensions

Choose the business process

- What your data cube is “about”
- Determined by the questions you want to answer about your organization

Question	Business Process
Who is my best customer?	Sales
What are my highest selling products?	Sales
Which teachers have the best student performance?	Standardized testing
Which supplier is offering us the best deals?	Purchasing

Note that a “business process” is not always about business.

Identify the fact

The fact table contains data associated with the business process event

Keys

- Primary key for each event
- Foreign keys for the associated dimensions
- Example: Sales has Sales_ID as primary key, and Product_ID, Store_ID, and Time_ID as foreign keys

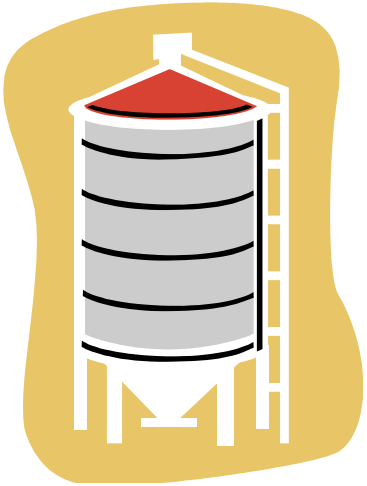
Measured, numeric data

- Facts: Quantifiable information for each business event
- Describes a particular combination of dimensional data
- Example: Sales has quantity_sold and total_price.

Try it for the "student performance" example.

Decide on the level of granularity

- Level of detail for each event (row in the table)
- Will determine the data in the dimensions



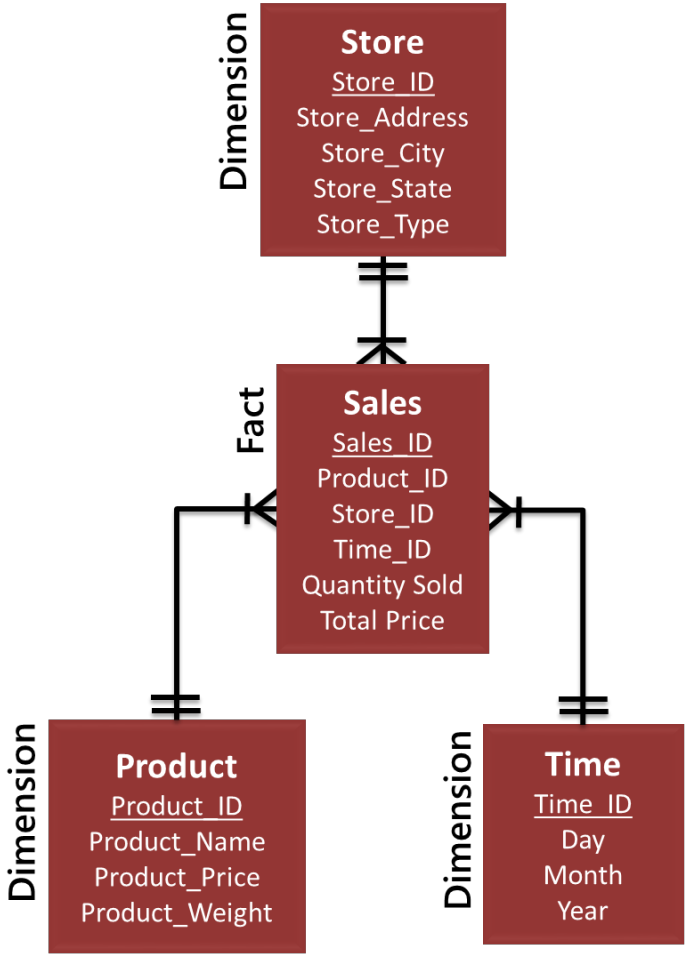
- Example: Who is my best customer?
  - The “event” is a sales transaction
  - Choices for time: yearly, quarterly, monthly, daily
  - Choices for store: store, city, state

How would you select the right granularity?

Identify the dimensions

- Description of the context of the business process
- The key elements of the process needed to answer the question ("fact")
  - who, what, where, when, why, and how

- Example: Sales transaction
  - A "sale" is the fact
  - Occurs for a particular product, store, and time
  - Could this data mart tell you
    - What is the best selling product?
    - Who is the best customer?



Try it for the "student performance" example.

Advantages of data cube

Speed

- Fast response to give you the information you have previously designed in the cube

Analysis

- The data multi-dimensional data structure allows the data to be analyzed in the most logical way.

Data cube caveats

- The cube is "**non volatile**," so you're locked in
  - Measured facts
  - Dimensions
  - Granularity

- So choose wisely!
  - For example: You can't track daily sales if "date" is monthly
  - So why not include every single sale and do no aggregation?

Pivot tables in Excel

- **PivotTable** is a data summarization tool in Excel
  - the easiest way to learn multidimensional data and generate simple reports

- Data cubes can act as the data source for Pivot Table in Excel

Summary

- Data warehouse vs. data mart vs. data cube
- Data Cube
- Star schema
- Kimball’s four step process for data mart design
- Pivot tables in Excel