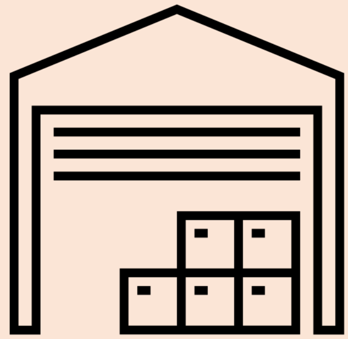
To cite this presentation: Pendyala, V.S. (2023) "Slowly changing dimensions and fast changing facts - the story of the traditional Datawarehouse". PPT Presentation. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Slowly changing dimensions and fast changing facts - the story of the traditional Datawarehouse



Vishnu S. Pendyala, PhD

Video Recording:
https://www.youtube.com/watch?v=BcttdNrbBhk

The beginnings

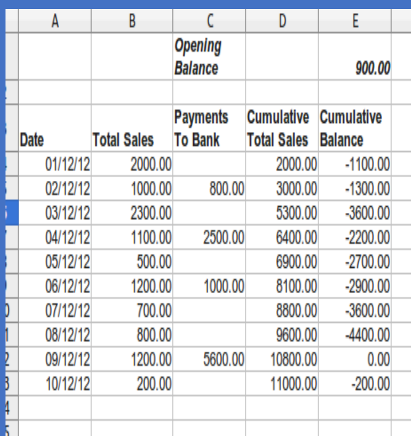
- Inmon, W. H. (1992). Building the Data Warehouse. Wiley.
- Codd, E. F., Codd, S. B., & Salley, C. T. (1993). Providing OLAP (on-line analytical processing) to user-analysts. *An IT Mandate. White Paper. Arbor Software Corporation*, *4*.
- Kimball, R. (1996). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (1st ed.). Wiley.

Example of a business insight:

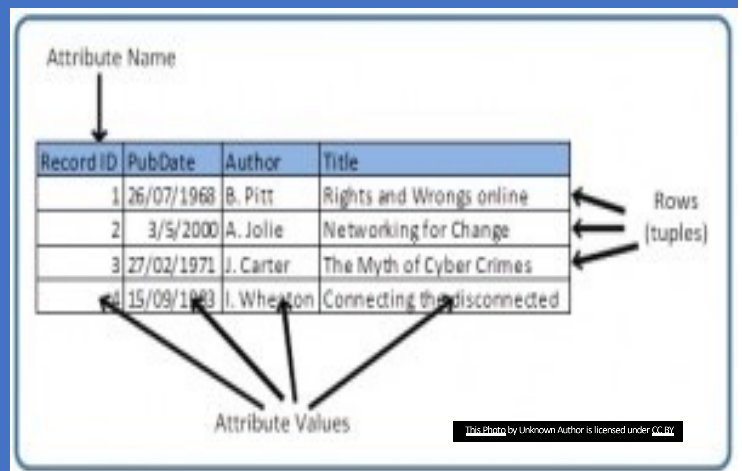
Company ABC made $234,567 from the sale of 13” MacBookPro to customer XYZ

How do we generate this?

Take 1: Generate insights from existing databases



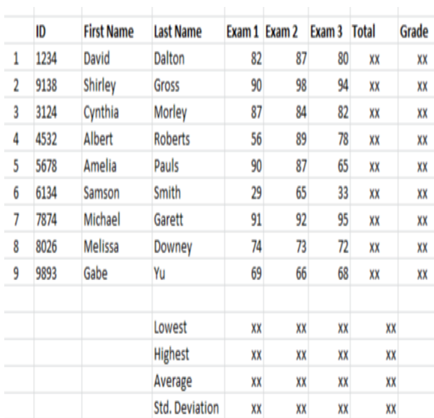
A	B	C	D	E
		Opening Balance		*900.00*
Date	Total Sales	Payments To Bank	Cumulative Total Sales	Cumulative Balance
01/12/12	2000.00		2000.00	-1100.00
02/12/12	1000.00	800.00	3000.00	-1300.00
03/12/12	2300.00		5300.00	-3600.00
04/12/12	1100.00	2500.00	6400.00	-2200.00
05/12/12	500.00		6900.00	-2700.00
06/12/12	1200.00	1000.00	8100.00	-2900.00
07/12/12	700.00		8800.00	-3600.00
08/12/12	800.00		9600.00	-4400.00
09/12/12	1200.00	5600.00	10800.00	0.00
10/12/12	200.00		11000.00	-200.00



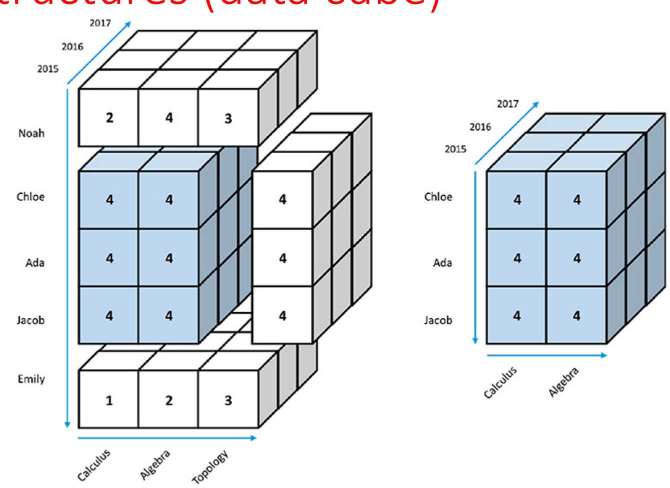
Record ID	PubDate	Author	Title
1	26/07/1968	B. Pitt	Rights and Wrongs online
2	3/5/2000	A. Jolie	Networking for Change
3	27/02/1971	J. Carter	The Myth of Cyber Crimes
4	15/09/1983	I. Wheaton	Connecting the disconnected

What is the problem?

Take 2: Generate the aggregates during the night when systems are not busy and store them in efficient data structures (data cube)

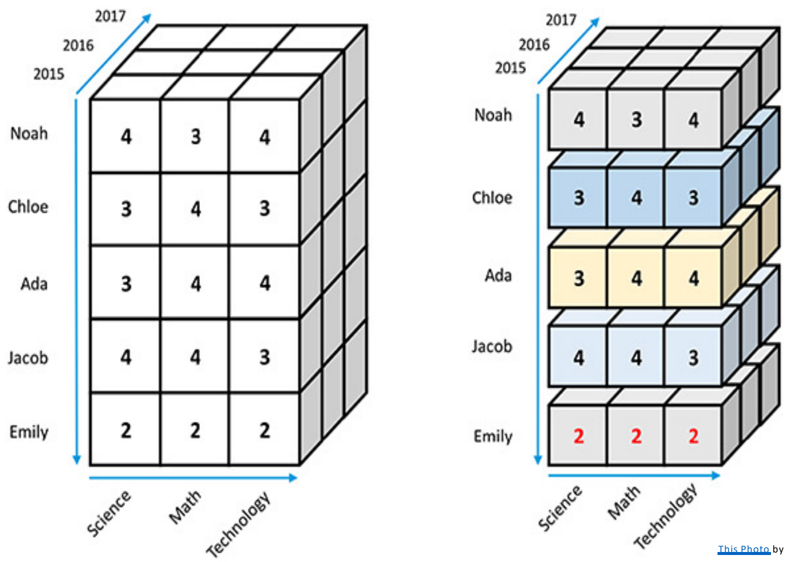


	ID	First Name	Last Name	Exam 1	Exam 2	Exam 3	Total	Grade
1	1234	David	Dalton	82	87	80	xx	xx
2	9138	Shirley	Gross	90	98	94	xx	xx
3	3124	Cynthia	Morley	87	84	82	xx	xx
4	4532	Albert	Roberts	56	89	78	xx	xx
5	5678	Amelia	Pauls	90	87	65	xx	xx
6	6134	Samson	Smith	29	65	33	xx	xx
7	7874	Michael	Garett	91	92	95	xx	xx
8	8026	Melissa	Downey	74	73	72	xx	xx
9	9893	Gabe	Yu	69	66	68	xx	xx
			Lowest	xx	xx	xx	xx	
			Highest	xx	xx	xx	xx	
			Average	xx	xx	xx	xx	
			Std. Deviation	xx	xx	xx	xx	

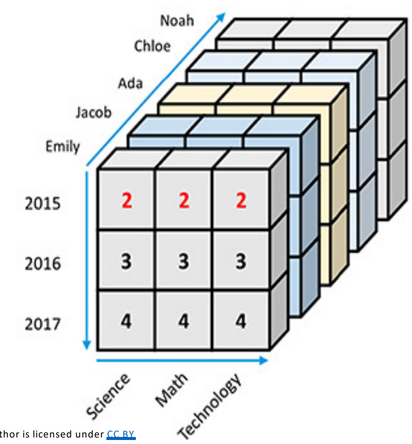


Dimensions of the cube: Subject, Student, Year

Numbers in the subcubes: GPA



Operations: Slice, Dice, Roll-up, pivot, etc



Fact
(fast changing, hopefully)

Business insight:

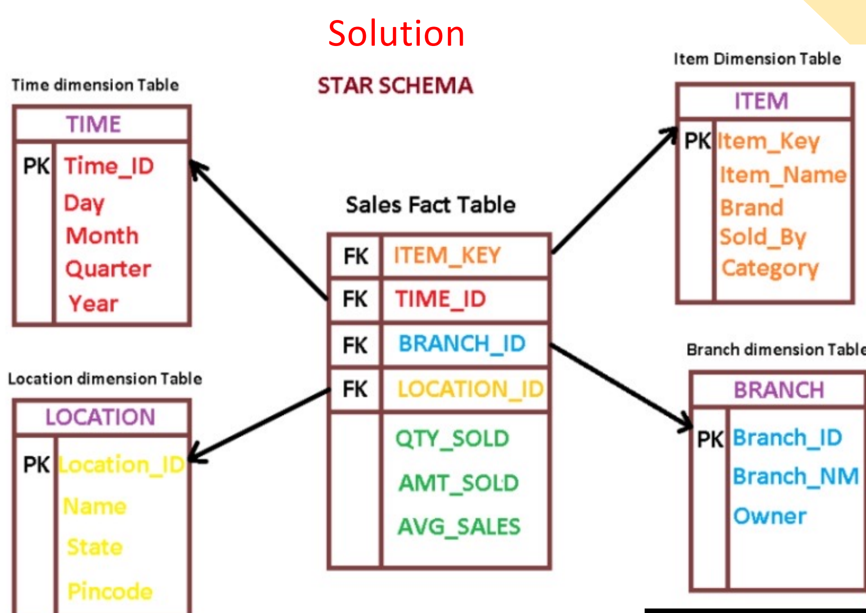
Company ABC made $234,567 from the sale of 13” MacBookPro to customer XYZ

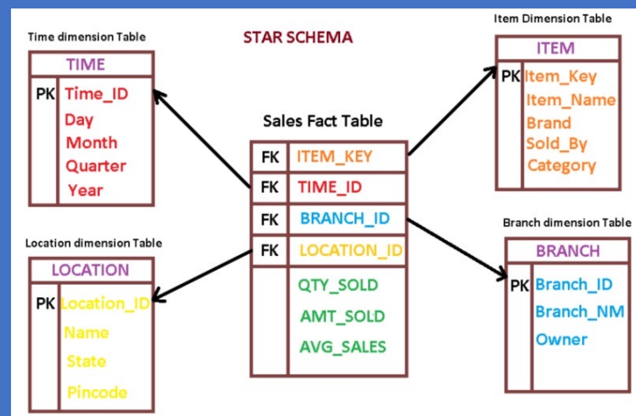
Dimension
(slowly changing, hopefully)

Dimension
(slowly changing, hopefully)

There's still a problem!

- Insights require data from many sources, some of which may not even be databases (files, spreadsheets, etc) => need to Extract, Transform, and Load
- Need to preserve historical data with a specific need to find temporal patterns
- Performance does not scale as analytical processing needs grow; transaction processing cannot take a hit
- Normalization can come in the way of efficient analytical query processing => need for a different way to model data
- => Cannot mix OLTP (operational) and OLAP (informational) systems!

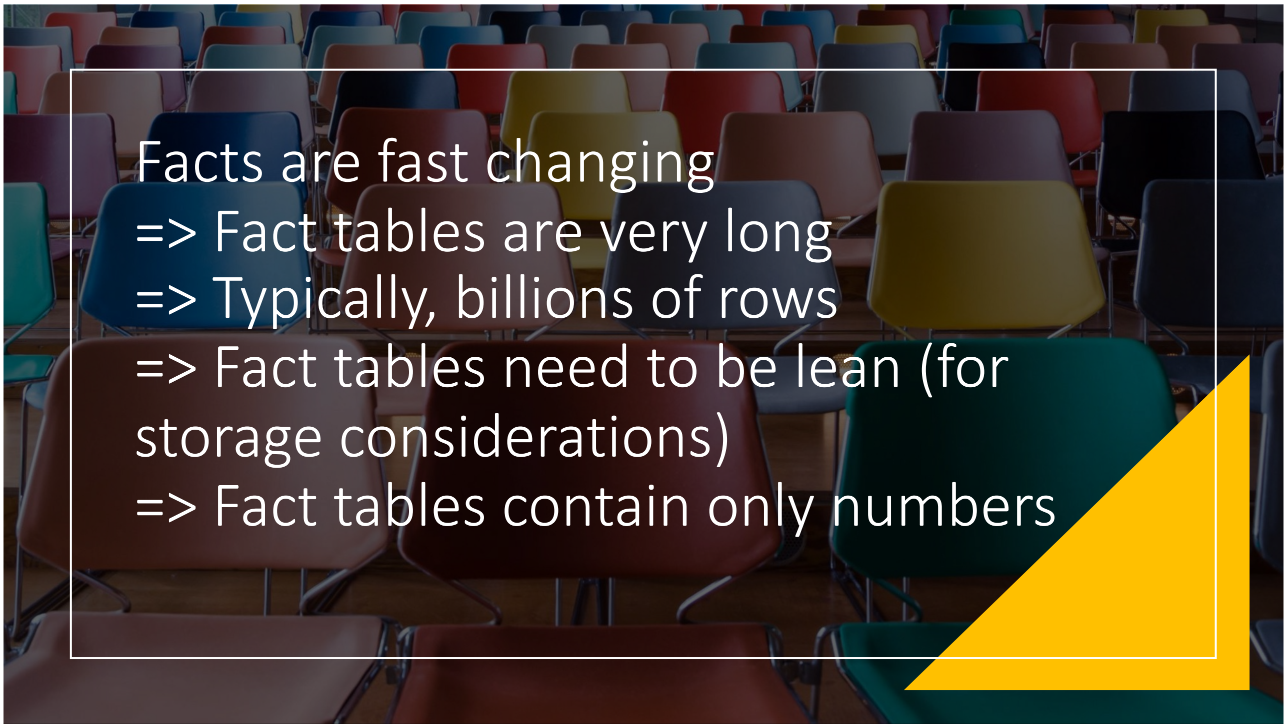


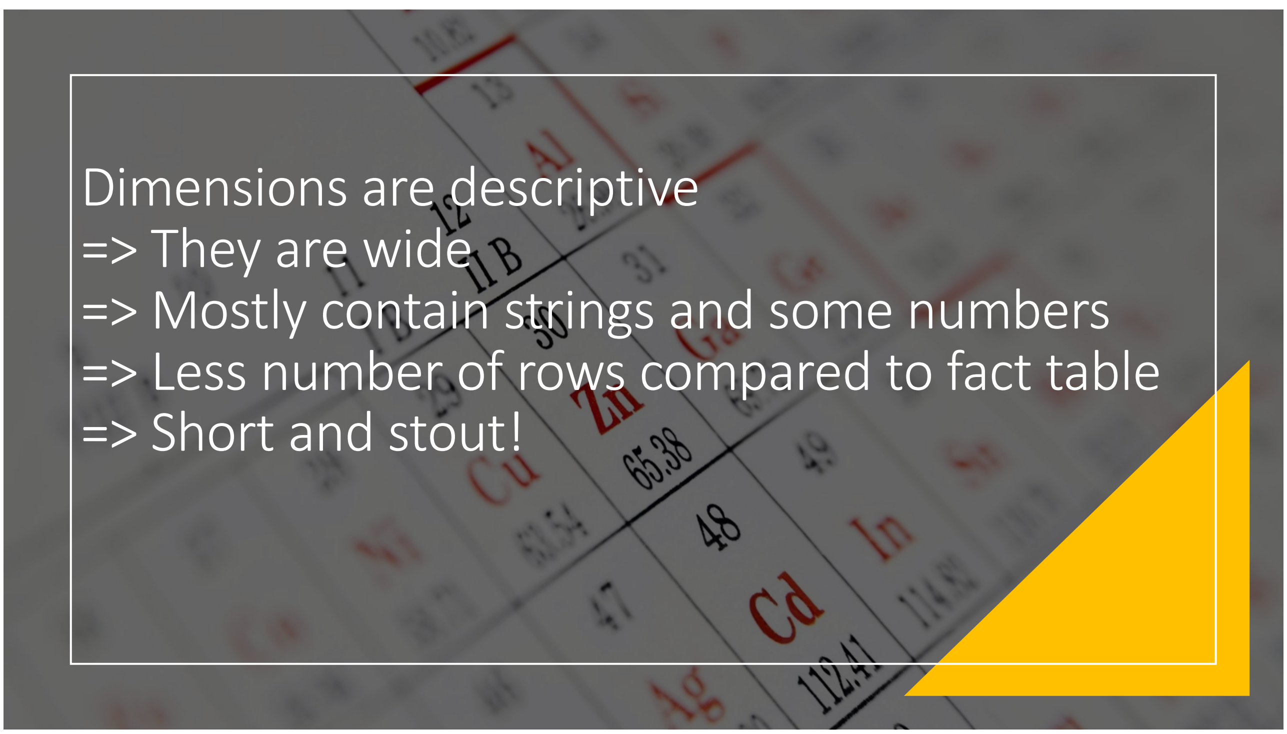


An Example

Fact Table: Sales

sale_id	product_id	time_id	location_id	branch_id	quantity_sold	amount
1	101	1	1	1	5	150
2	102	2	2	2	3	90
3	101	3	1	1	2	60
4	103	4	3	3	7	210
5	102	5	2	2	4	120


Facts are fast changing
=> Fact tables are very long
=> Typically, billions of rows
=> Fact tables need to be lean (for storage considerations)
=> Fact tables contain only numbers


Dimensions are descriptive
=> They are wide
=> Mostly contain strings and some numbers
=> Less number of rows compared to fact table
=> Short and stout!

Dimension Table: Time

time_id	date	day_name
1	2023-05-10	Monday
2	2023-05-11	Tuesday
3	2023-05-12	Wednesday
4	2023-05-13	Thursday
5	2023-05-14	Friday

Dimension Table: Location

location_id	location_name
1	City A
2	City B

Dimension Table: Product

product_id	product_name	category_id
101	Laptop	1
102	Smartphone	2
103	Tablet	1

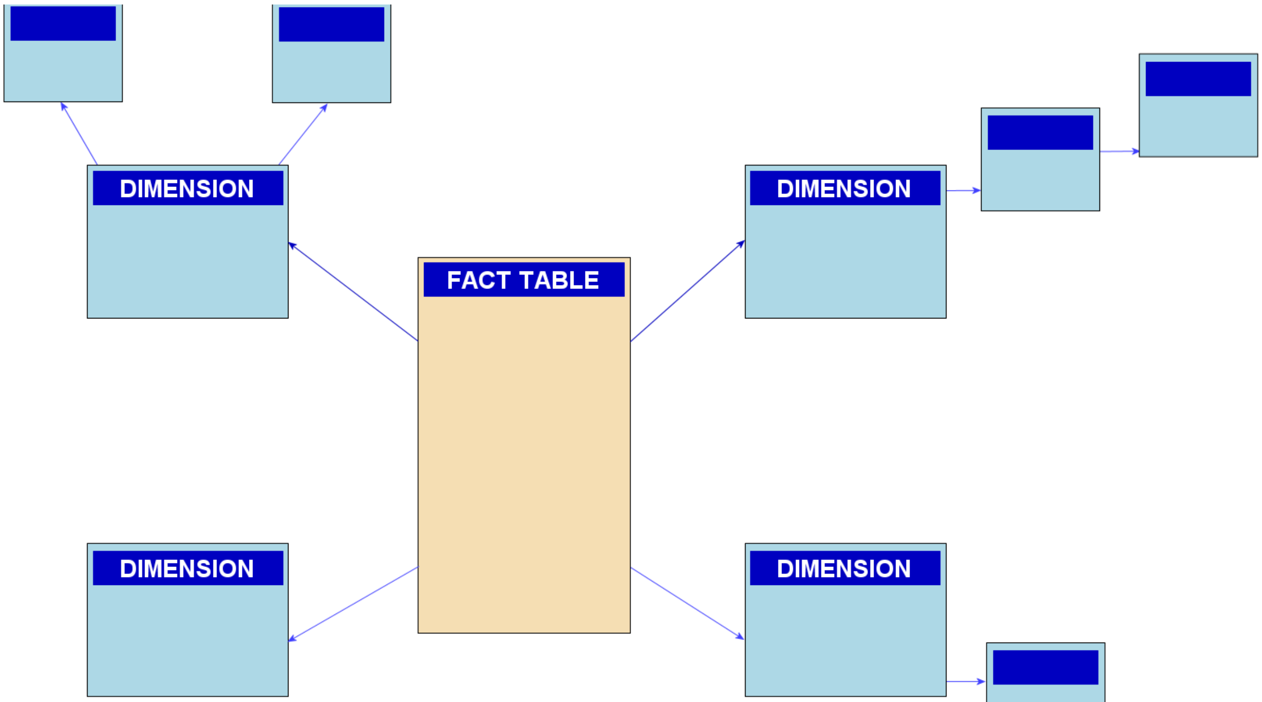
Dimension Table: Branch

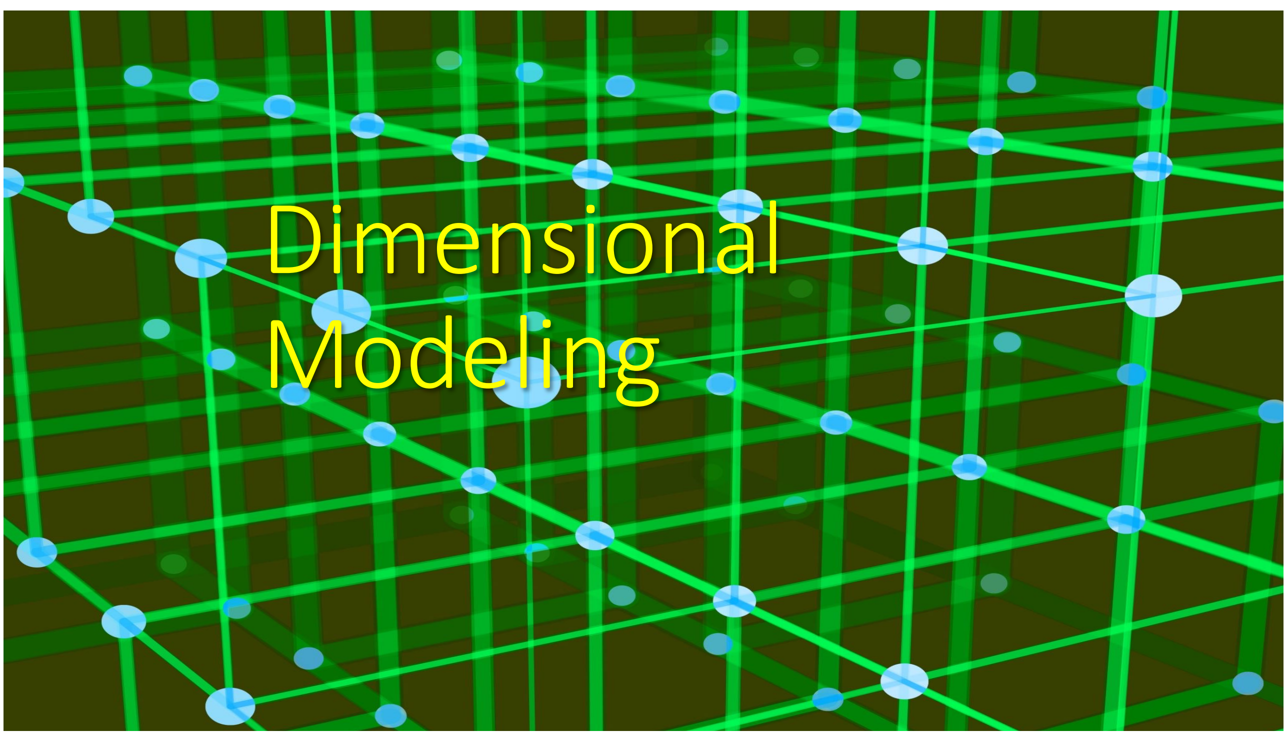
branch_id	branch_name
1	Store A
2	Store B
3	Store C

What are the total sales amount for each product in City A during the month of May 2023?

```
SELECT p.product_name, SUM(s.amount) AS
total_sales_amount
FROM Sales s
JOIN Product p ON s.product_id = p.product_id
JOIN Location l ON s.location_id = l.location_id
JOIN Time t ON s.time_id = t.time_id
WHERE l.location_name = 'City A'
  AND t.date BETWEEN '2023-05-01' AND '2023-05-31'
GROUP BY p.product_name;
```

Snowflake schema


DIMENSION
DIMENSION
FACT TABLE
DIMENSION
DIMENSION


Dimensional Modeling

Key considerations

What are the dimensions and how many of them?	I/O bound queries => minimize storage
What numeric quantities will be stored in the Fact table?	
At what granularity will be facts be captured?	Tradeoff: finer grain => better detail for drilling down, but more rows

An example

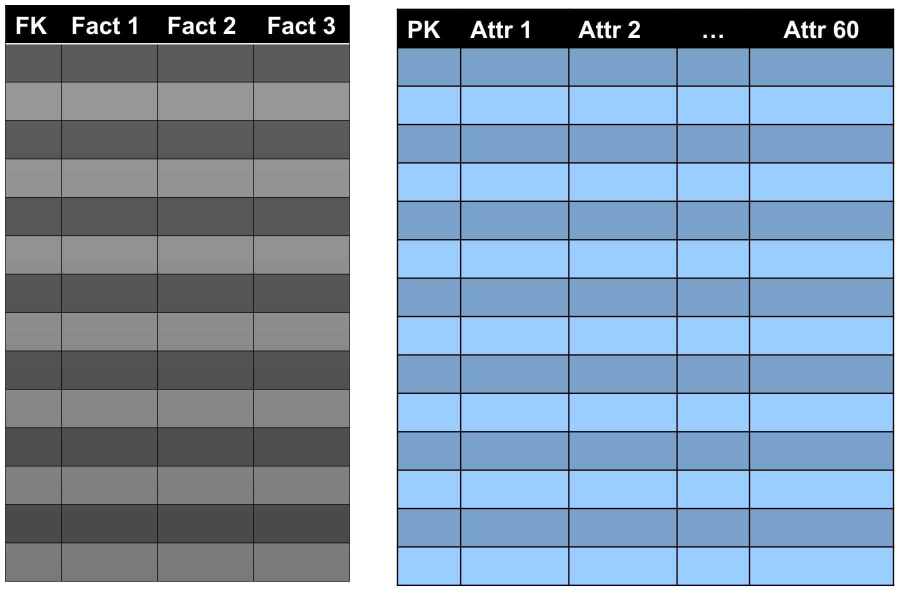
Suppose there are 60 dimensional attributes and 3 facts to be captured in the fact table

→

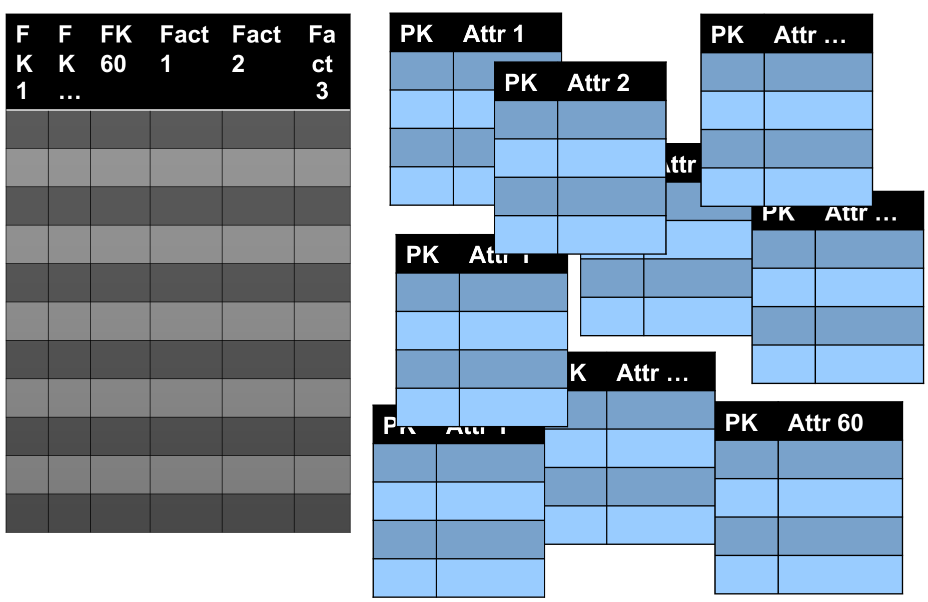
How do we design the star schema? How many dimension tables? Choices:

A. One dimension with all 60 attributes
B. Each attribute gets its own dimension table
C. Four dimensions

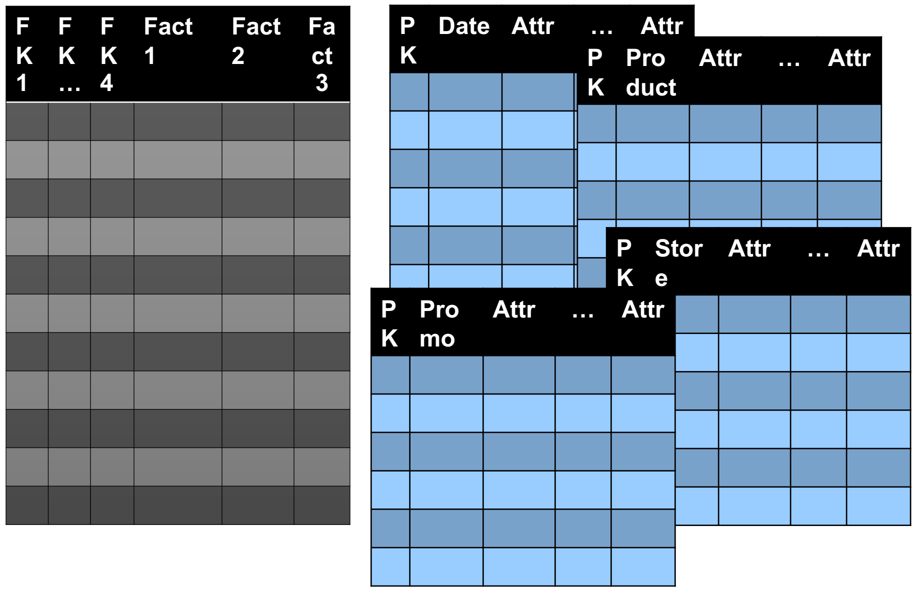
A: One dimension with all 60 attributes



B: Each attribute gets its own dimension table



C: Four dimensions



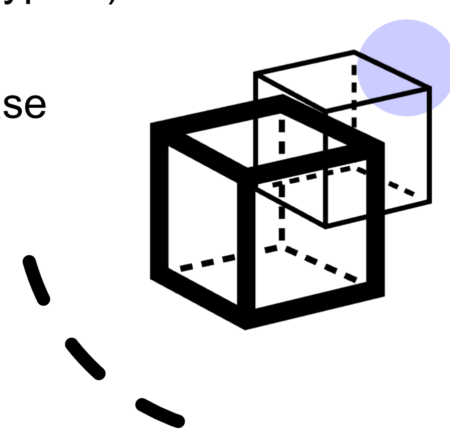
- Option A is the worst – dimension can have as many rows as the fact table!
- Option B – dimensions occupy less storage, but the fact tables will be very wide and occupy huge storage, still less than in Option A
- Option C is the best – do the math!

Slowly Changing Dimensions

SCD Type 1, 2, and 3
and also
Type 0, 4, 5, …

How do we handle rarely occurring changes to dimensions?

- Do nothing – ignore the changes (type 0)
- Overwrite the existing values – erase history (type 1)!
- Find ways to preserve the history
  - Type 2, 3, 4, …



SCD Type 1: In-Place Update

Initial Customer Dimension

customer_id	customer_name	city	loyalty_status
101	John Smith	New York	Gold
102	Jane Doe	Los Angeles	Silver

Requested Customer update

customer_id	customer_name	city	loyalty_status
101	John Smith	Boston	Platinum

Updated Customer Dimension

customer_id	customer_name	city	loyalty_status
101	John Smith	Boston	Platinum
102	Jane Doe	Los Angeles	Silver

SCD Type 2: Historical Tracking

Initial Customer Dimension

customer_id	customer_name	city	loyalty_status
101	John Smith	New York	Gold
102	Jane Doe	Los Angeles	Silver

Requested Customer update

customer_id	customer_name	city	loyalty_status
101	John Smith	New York	Platinum

Updated Customer Dimension

customer_id	customer_name	city	loyalty_status	effective_date	end_date
101	John Smith	New York	Gold	2023-01-01	2023-08-10
101	John Smith	New York	Platinum	2023-08-11	(null)
102	Jane Doe	Los Angeles	Silver	(null)	(null)

SCD Type 3: Alternate Reality

Initial Customer Dimension

customer_id	customer_name	city	loyalty_status
101	John Smith	New York	Gold
102	Jane Doe	Los Angeles	Silver

Requested Customer update

customer_id	customer_name	city	loyalty_status
101	John Smith	New York	Platinum

Updated Customer Dimension

customer_id	customer_name	city	loyalty_status	prev_loyalty_status
101	John Smith	New York	Platinum	Gold
102	Jane Doe	Los Angeles	Silver	(null)

SCD Type 4: Monster / Mini Dimension

Initial Customer Dimension

customer_id	customer_name	city	loyalty_status
101	John Smith	New York	Gold
102	Jane Doe	Los Angeles	Silver

Loyalty_status is changing rapidly => create a new mini dimension for it

customer_id	customer_name	city
101	John Smith	New York
102	Jane Doe	Los Angeles

+

customer_id	loyalty_status	start_date	end_date
101	Gold	2023-01-01	2023-08-10

Updated Mini Dimension

customer_id	loyalty_status	start_date	end_date
101	Gold	2023-01-01	2023-08-10
101	Platinum	2023-08-11	(null)

More SCD types

- Hybrid approaches:
  - Type 5: Add mini-dimension (Type 4) and Type 1
  - Type 6: Type 1 + Type 2 + Type 3
  -