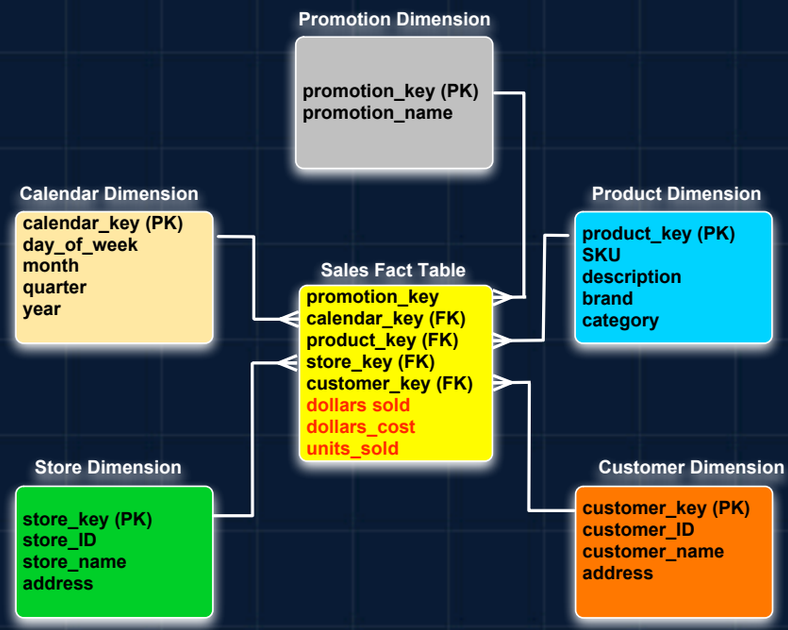
Concepts: The Fact Table



Eric Tremblay - Oracle Specialist
www.data-warehouse.ca
eric.tremblay@data-warehouse.ca

June 2011

Objective

Objective

Describe what is a Data-Warehouse Fact Table.

Concept, design and use.

Definition

Definition

«A fact table is a table with measures (column where you will perform an aggregation sum, …). In a classic star schema any table that only has many-to-one (N:1) joins to it is a fact table. »

Dimension Table

Dimension Table

- Simple primary key
- Textual attributes rich and adapted to the user
- Hierarchical reports
- Few codes; Few codes; codes should be decoded according to their descriptions
- Relatively small

Dimensional Modeling - Grain

Dimensional Modeling - Grain

The grain describe the level of detail.

Example of grain:

An individual line item on a customer's retail sales invoice.

A line item on a bill received from a doctor.

An individual boarding pass to get on a flight.

A daily snapshot of the inventory levels for each product in a warehouse.

A monthly snapshot for each bank account.

Hierarchy

Hierarchy

Hierarchy defines parent-child relationships among various levels within a single dimension.

For instance in a time dimension, year level is parent of four quarters, each of which is a parent of three months, which are parents of 28 to 31 days, which are parents of 24 hours. Similarly in a geography dimension a continent is a parent of countries, country could be a parent of states, and state could be a parent of cities.

Surrogate key

Surrogate key

Definition: Surrogate (1) – Hall, Owlett and Codd (1976)
A surrogate represents an entity in the outside world. The surrogate is internally generated by the system but is nevertheless visible to the user or application.

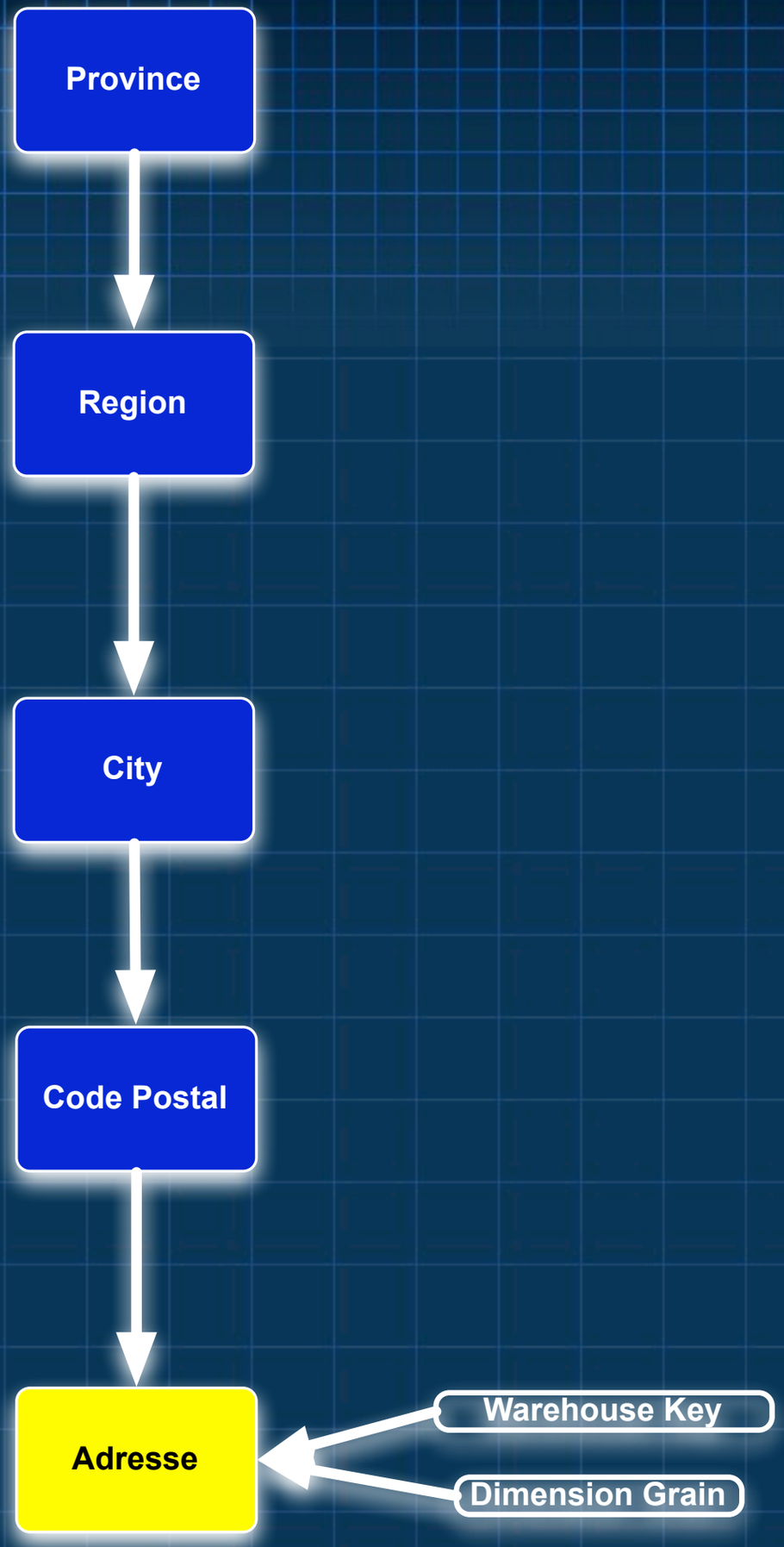
A surrogate key in a database is a unique identifier for either an entity in the modeled world or an object in the database. The surrogate key is not derived from application data.

Surrogate key can be the primary key, generated by the database management system and not derived from any application data in the database. Surrogate key is frequently a sequential number.

- The value is unique system-wide, hence never reused
- The value is system generated
- The value is not manipulable by the user or application
- The value contains no semantic meaning
- The value is not visible to the user or application
- The value is not composed of several values from different domains.

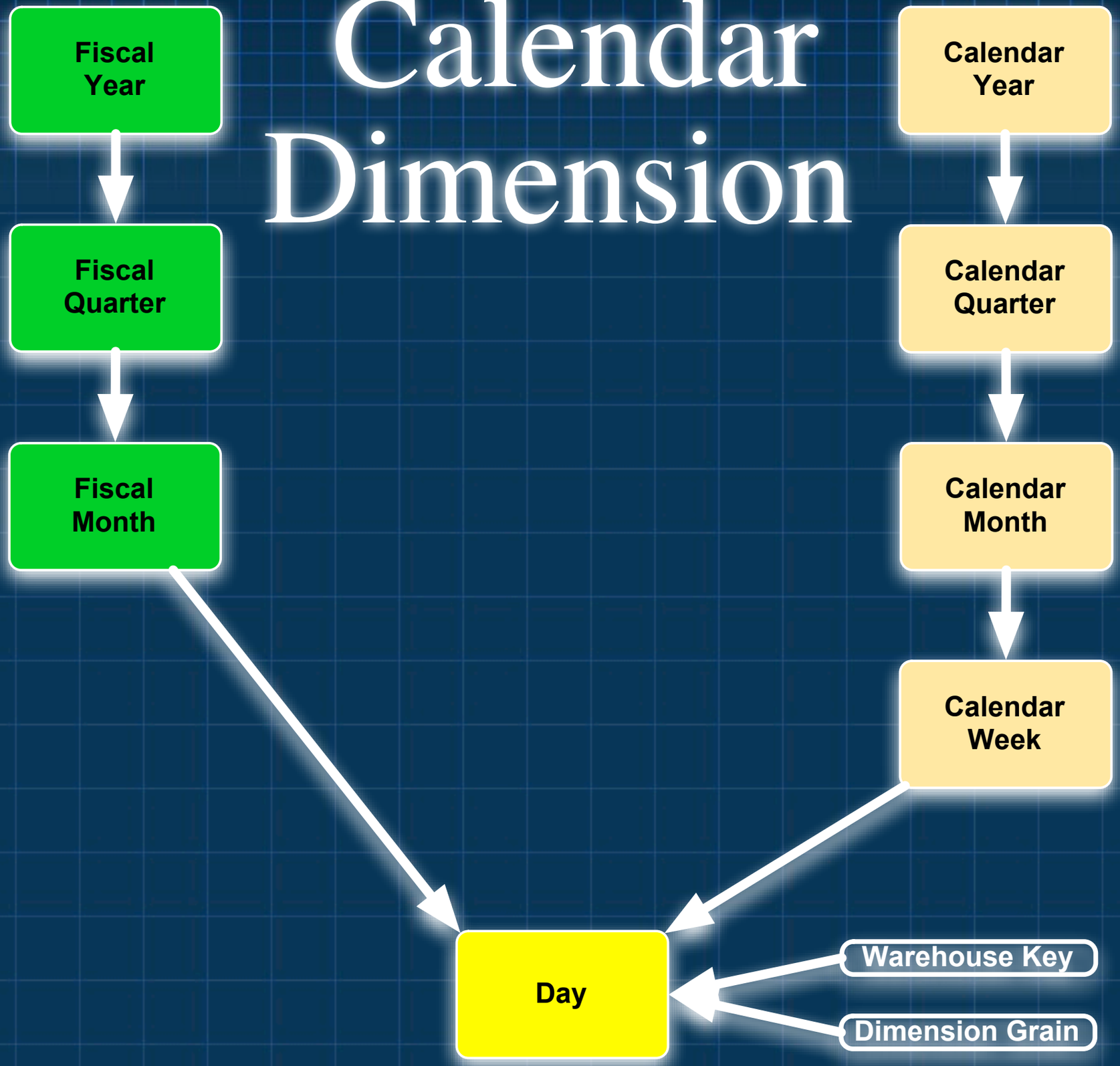
Dimension
Client
Location

Dimension Client Location



Calendar Dimension

Calendar Dimension



Slowly Changing Dimensions

Slowly Changing Dimensions

Type 1: **Overwrite changed attribute.**
A fact is associated with only the current value of a dimension column.

Slowly Changing Dimensions

Type 1: **Overwrite changed attribute.**
A fact is associated with only the current value of a dimension column.

Type 2: **Add new dimension record.**
A fact is associated with only the original value of a dimension column.

Slowly Changing Dimensions

Type 1: **Overwrite changed attribute.**
A fact is associated with only the current value of a dimension column.

Type 2: **Add new dimension record.**
A fact is associated with only the original value of a dimension column.

Type 3: **Use field for ‘old’ value.**
A fact is associated with both the original value and with the current value of a dimension column.

Slowly Changing Dimensions

Scenario: Customer last name changes from Pharand to Smith: Update Cust.Lname

Slowly Changing Dimensions

Scenario: Customer last name changes from Pharand to Smith: Update Cust.Lname

Original dimension table			
Cust Key	Cust ID	Cust Lname	Cust Fname
1	TE123	Tremblay	Eric
2	PJ456	Pharand	Josée

Slowly Changing Dimensions

Scenario: Customer last name changes from Pharand to Smith: Update Cust.Lname

Original dimension table			
Cust Key	Cust ID	Cust Lname	Cust Fname
1	TE123	Tremblay	Eric
2	PJ456	Pharand	Josée

Slowly changing dimension Type 1			
Cust Key	Cust ID	Cust Lname	Cust Fname
1	TE123	Tremblay	Eric
2	PJ456	Smith	Josée

Slowly Changing Dimensions

Slowly Changing Dimensions

Slowly changing dimension Type 2						
Cust Key	Cust ID	Cust Lname	Cust Fname	Row Current?	Row Start	Row Stop
1	TE123	Tremblay	Eric	yes	12/3/99	
2	PJ456	Pharand	Josée	no	2/20/00	9/15/00
3	PJ789	Smith	Josée	yes	9/16/00	

Slowly Changing Dimensions

Slowly changing dimension Type 2						
Cust Key	Cust ID	Cust Lname	Cust Fname	Row Current?	Row Start	Row Stop
1	TE123	Tremblay	Eric	yes	12/3/99	
2	PJ456	Pharand	Josée	no	2/20/00	9/15/00
3	PJ789	Smith	Josée	yes	9/16/00	

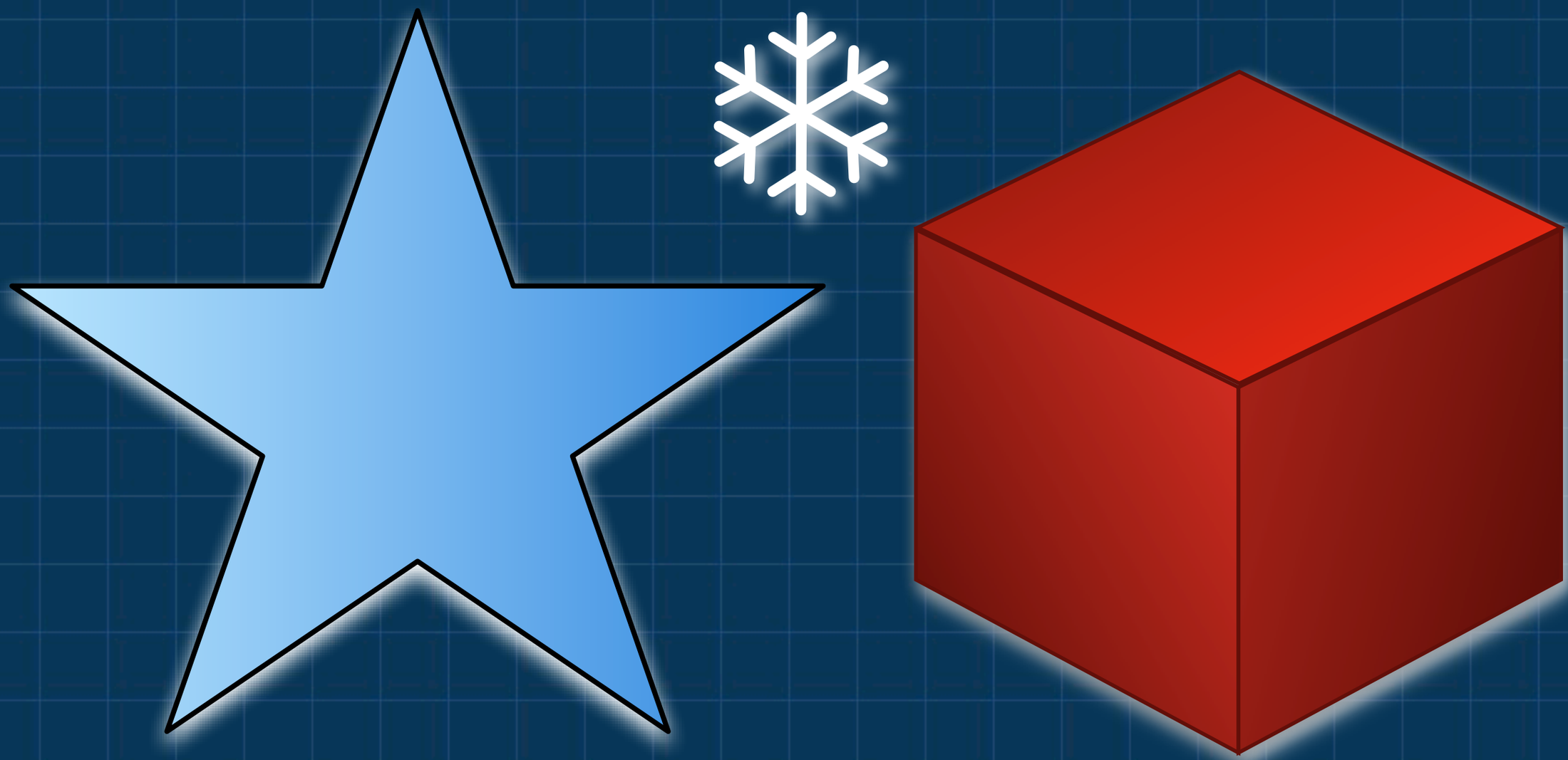
Slowly changing dimension Type 3				
Cust Key	Cust ID	Cust Lname	Cust Fname	Old Value
1	TE123	Tremblay	Eric	
2	PJ456	Smith	Josée	Pharand

Fact Table

Star & Snowflake Schema

Fact Table

Star & Snowflake Schema



Fact Table

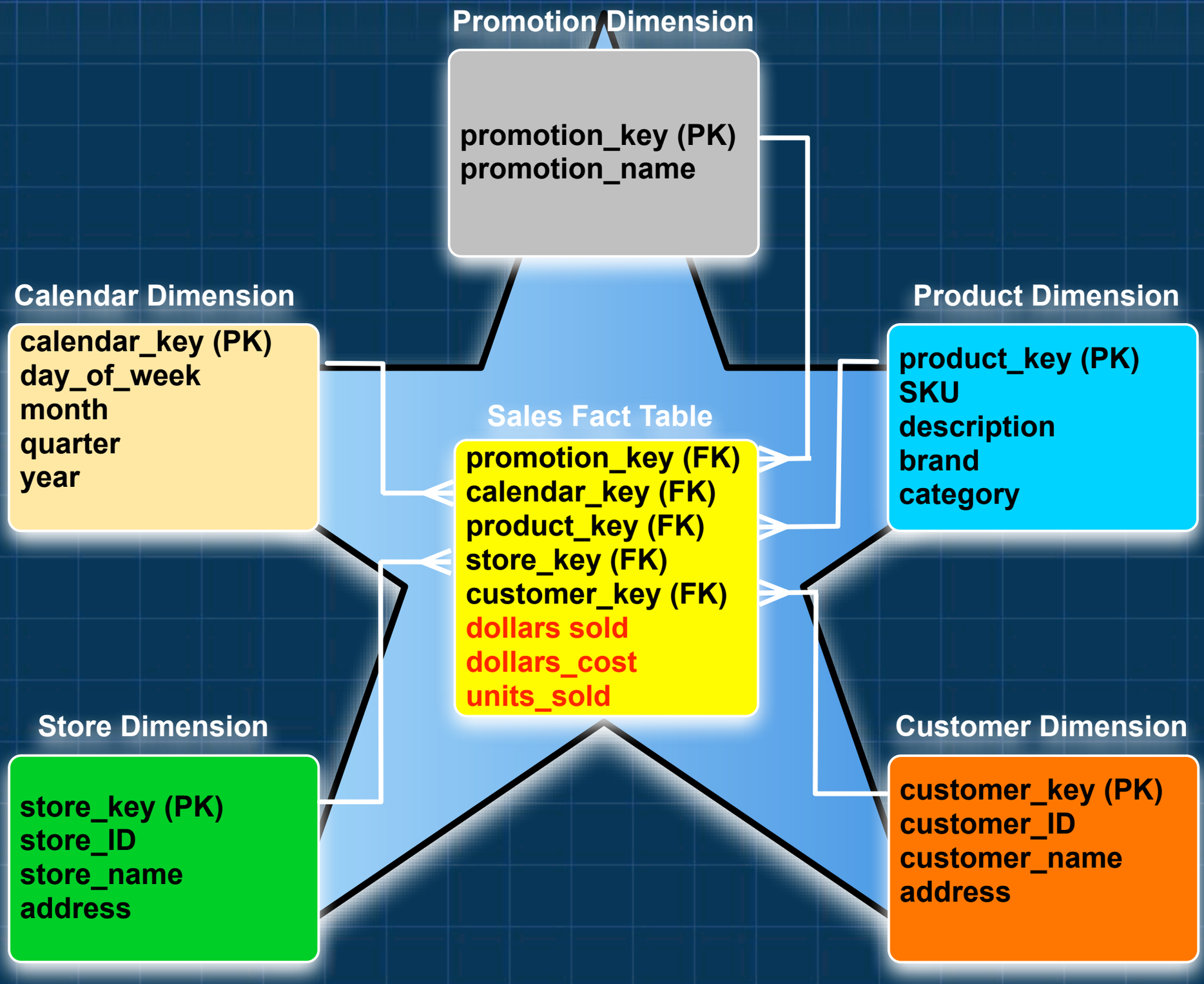
- A fact table is a table with measures. Measures need to be defined in a logical fact; any column with an aggregation rule is a measure.
- Composed Primary key
- The date/time is almost always a key
- The facts are usually numerical
- The facts are in general additive

Fact Table

Star Schema

Fact Table

Star Schema



Fact Table

Star Schema

Which product has sold the most?

Promotion Dimension

promotion_key (PK)
promotion_name

Calendar Dimension

calendar_key (PK)
day_of_week
month
quarter
year

Product Dimension

product_key (PK)
SKU
description
brand
category

Sales Fact Table

promotion_key (FK)
calendar_key (FK)
product_key (FK)
store_key (FK)
customer_key (FK)
dollars sold
dollars_cost
units_sold

Store Dimension

store_key (PK)
store_ID
store_name
address

Customer Dimension

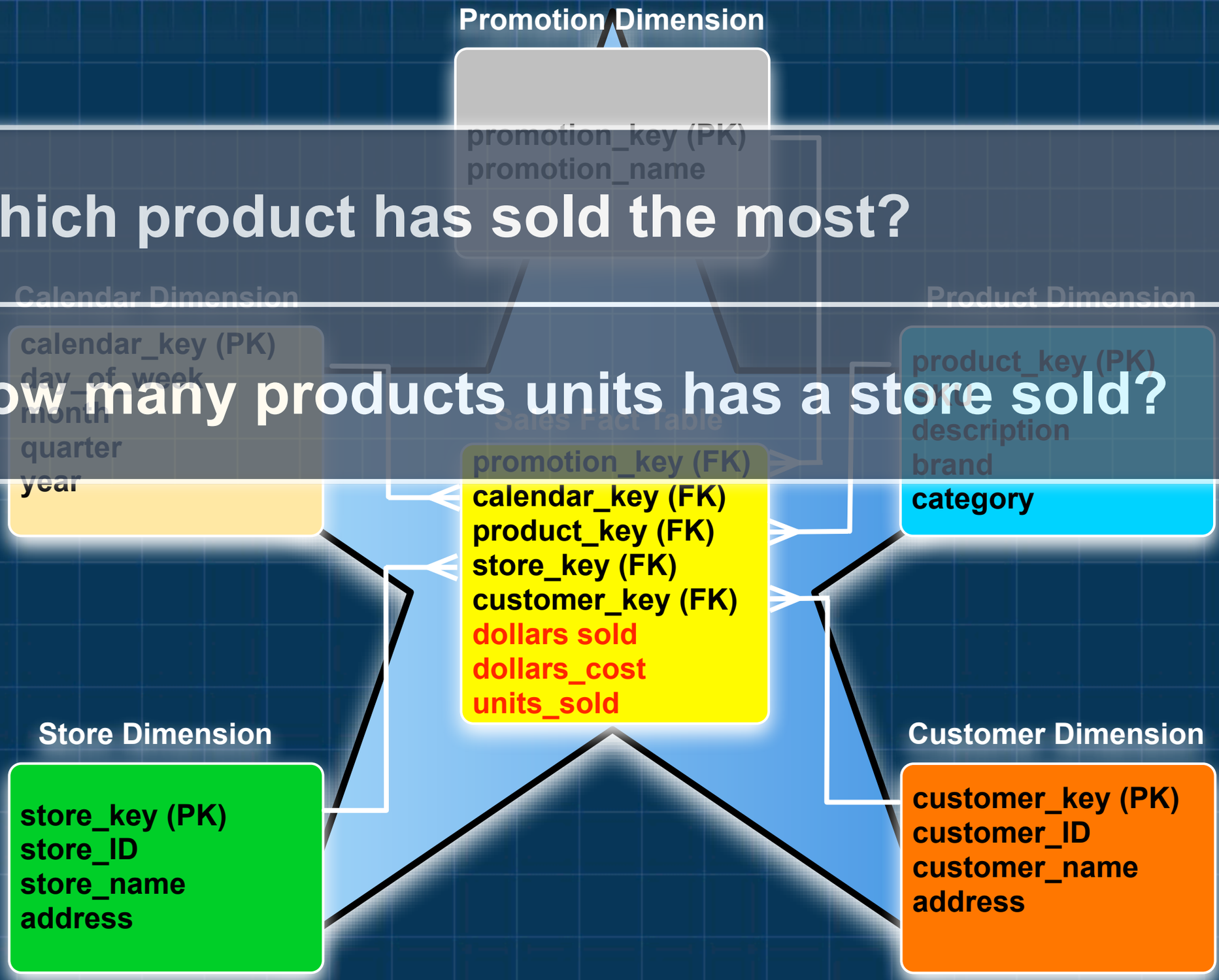
customer_key (PK)
customer_ID
customer_name
address

Fact Table

Star Schema

Which product has sold the most?

How many products units has a store sold?



Fact Table

Star Schema

Promotion Dimension

promotion_key (PK)
promotion_name

Which product has sold the most?

Calendar Dimension

calendar_key (PK)
day_of_week
month
quarter
year

Product Dimension

product_key (PK)
sku
description
brand
category

How many products units has a store sold?

Sales Fact Table

promotion_key (FK)
calendar_key (FK)
product_key (FK)
store_key (FK)
customer_key (FK)
dollars_sold
dollars_cost
units_sold

What is the best month of sales?

Store Dimension

store_key (PK)
store_ID
store_name
address

Customer Dimension

customer_key (PK)
customer_ID
customer_name
address

Fact Table

Star Schema

Promotion Dimension

promotion_key (PK)
promotion_name

Which product has sold the most?

Calendar Dimension

calendar_key (PK)
day_of_week
month
quarter
year

Product Dimension

product_key (PK)
sku
description
brand
category

How many products units has a store sold?

Sales Fact Table

promotion_key (FK)
calendar_key (FK)
product_key (FK)
store_key (FK)
customer_key (FK)
dollars sold
dollars_cost
units_sold

What is the best month of sales?

Store Dimension

store_key (PK)
store_ID
store_name
address

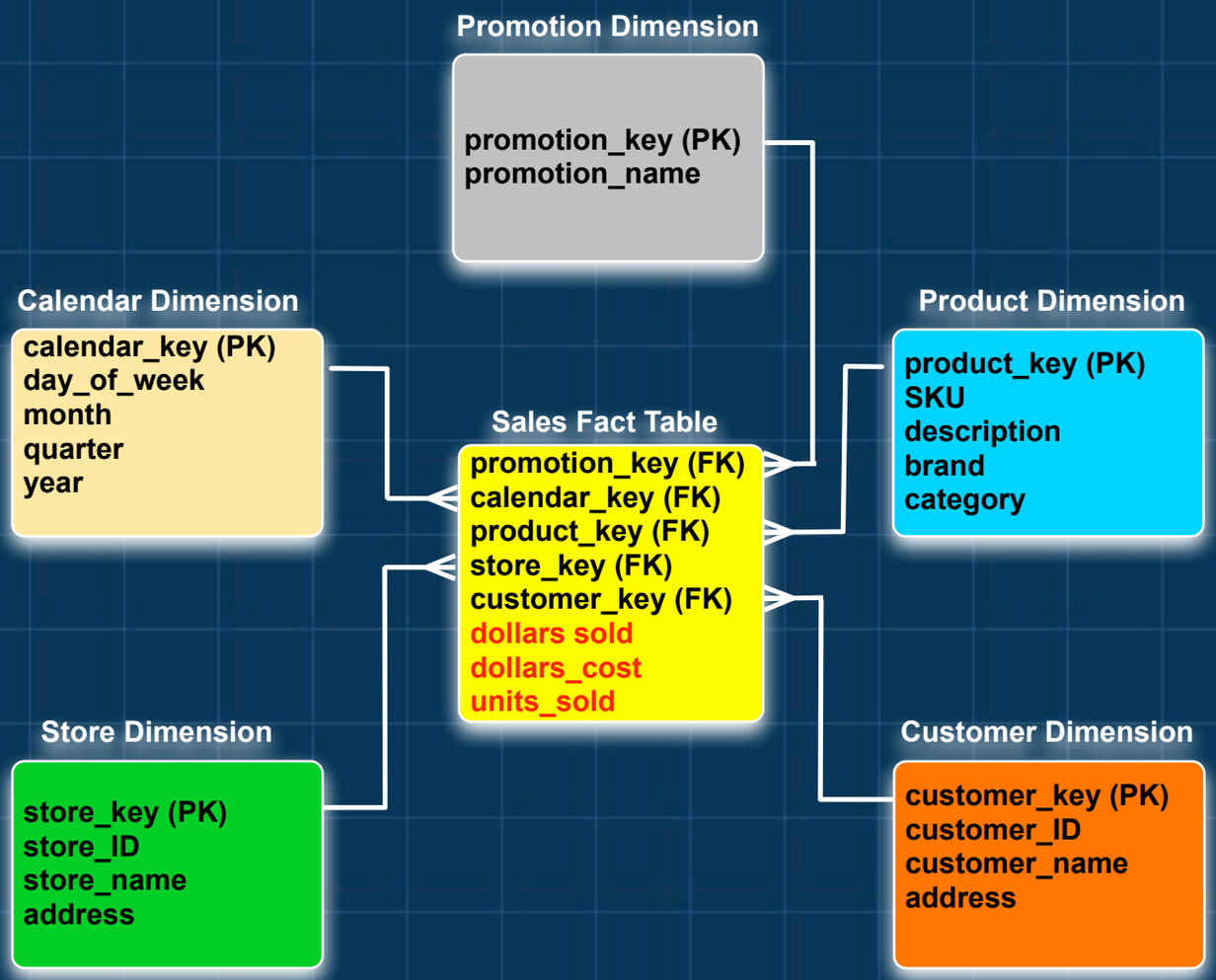
Customer Dimension

customer_key (PK)
customer_ID
customer_name
address

What promotion has worked best?

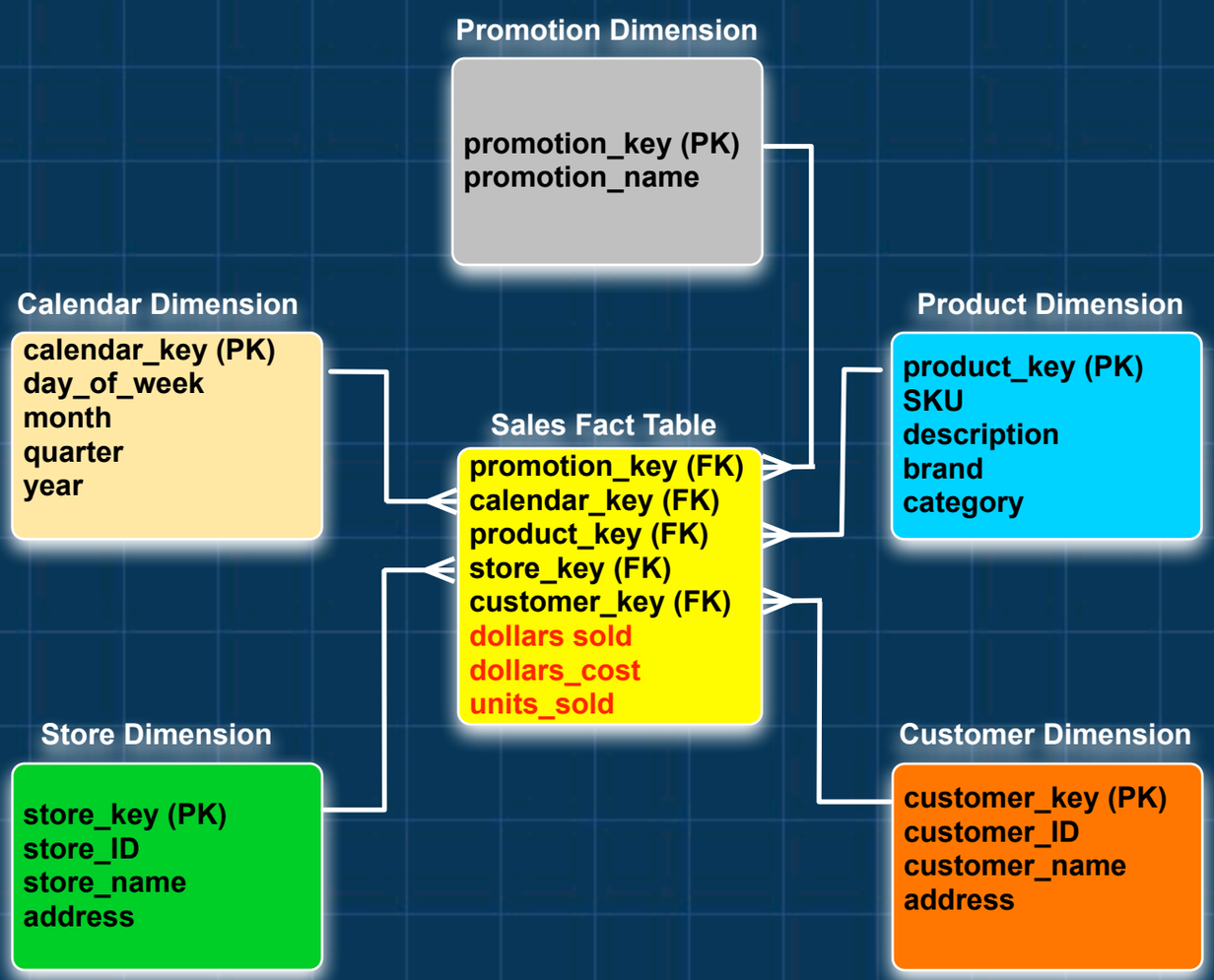
Loading Transactions

One fact = One business process



Loading Transactions

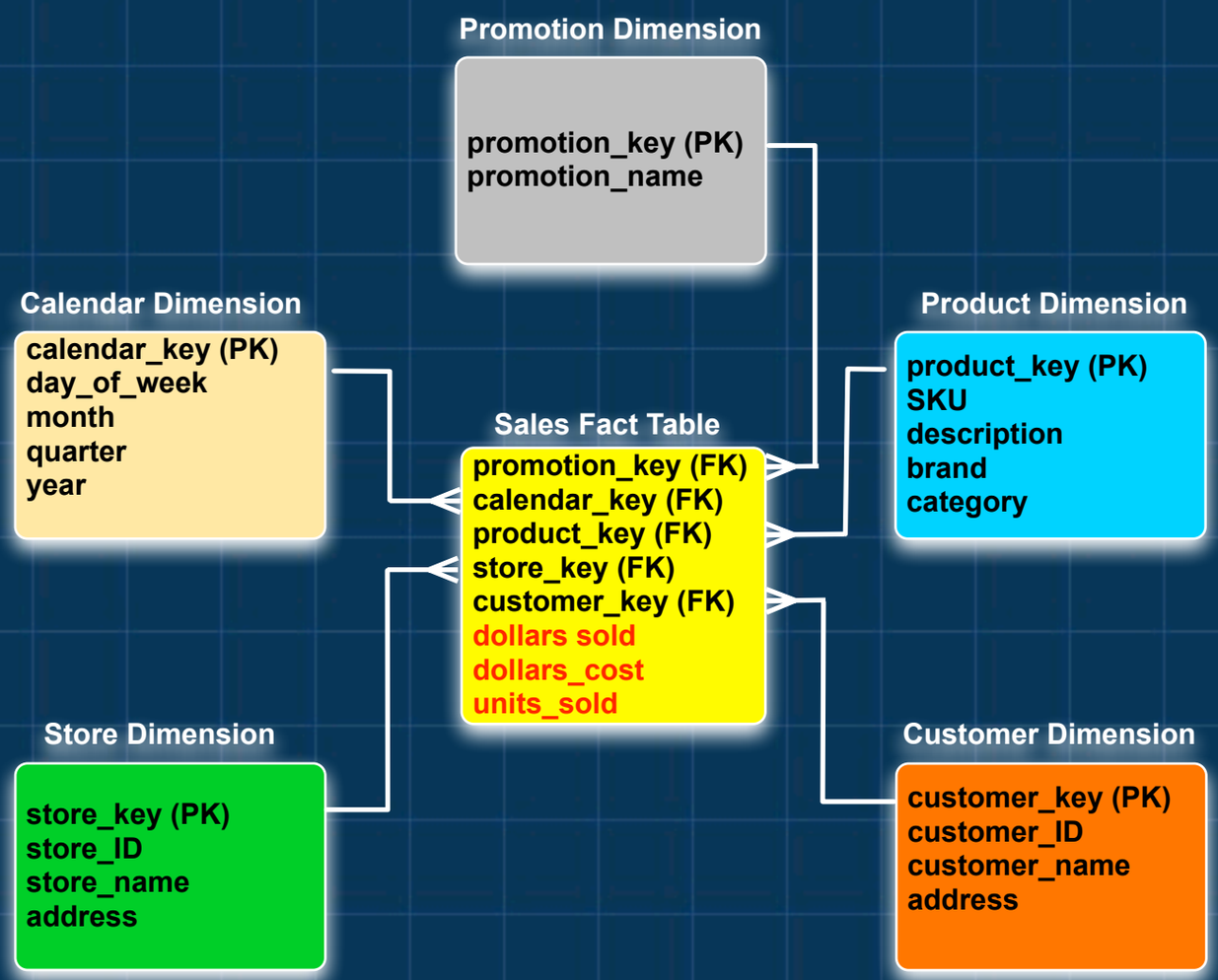
One fact = One business process



Sale 1: Feb 10, 2011, Store #19, Customer #5, Product #2259, Promotion #35, Price $20, Cost $10, Units Sold 1

Loading Transactions

One fact = One business process



Sale 1: Feb 10, 2011, Store #19, Customer #5, Product #2259, Promotion #35, Price $20, Cost $10, Units Sold 1

Sale 2: Feb 10, 2011, Store #5, Customer #118, Product #5682, Promotion #23, Price $10, Cost $5, Units Sold 5

Measures

Measures

A measure is a property on which calculations (e.g., sum, count, average, minimum, maximum) can be made.

- **Additive measures** are measures that can be added across all dimensions. For example dollars of sales can be added across all dimensions within a retail store warehouse.

- **Semi-additive measures** are measures that can be added across some, but not all dimensions. For example the bank account balance is simply a snapshot in time and cannot be summed over time. However you could add multiple accounts of the same customer to get the total balance for that customer.

- **Non-additive measures** are measures that cannot be added across any dimensions. For example the inventory is simply a snapshot in time and cannot be summed over time. Nor can you combine inventory for various products.

Level

Level

Level is a column within a dimension table that could be used for aggregating data. For example, product dimension could have levels of product type (beverage), product category (alcoholic beverage), product class (beer), product name (miller lite, budlite, corona, etc).

Member

Member

Member is a value within a dimension level that can be used for aggregating and reporting data. For example each product category such as beverage, non-consumable, food, clothing, etc is a member. Each product class such as beer, wine, coke, bottled water would represent a member.

Data Mart

Data Mart

Data Mart is a subset of the data warehouse typically serving a functional area such as marketing or finance, or particular location of the business (for instance mid-Western division).

Fact Table

Snowflake Schema

Calendar Dimension

calendar_key (PK)
month_key (FK)
year

month_key (PK)
year

year_key (PK)

Product Dimension

product_key (PK)
SKU
description
brand_key (FK)
category

brand_key (PK)
brand_description

Sales Fact Table

calendar_key (FK)
product_key (FK)
store_key (FK)
customer_key (FK)
dollars_sold
dollars_cost
units_sold

Store Dimension

store_key (PK)
store_ID
store_nom
address

Customer Dimension

customer_key (PK)
customer_ID
customer_name
address

Normalisation

Normalisation

Simplicity

Normalisation

Simplicity

Query Performance

Normalisation

Simplicity

Query Performance

Minor disk space saving

Normalisation

Simplicity

Query Performance

Minor disk space saving

Slows down the users' ability to browse

The Difference

Data Base	Data Warehouse
Actual	Historical
Internal	Internal and External
Isolated	Integrated
Transactions	Analysis
Normalised	Dimensional
Dirty	Clean and Consistent
Detailed	Detailed and Summary

Data Warehouse

End