Contents

Tips, Tricks, and Pitfalls

Level 3

I: Scripty Toyz

For all of the examples in this document we are going to look at a fictional toy store, **Scripty Toyz**.

At Scripty Toyz, our job as CTO is to manage all of the data flow, no matter if it's sales or inventory.

Here are the following tables we will build for our newfound firm:

Table Name	Purpose	Type
SALES_HIST	History of customer sales	FACT
PRODUCT_CATEGORY	Bucket products into different groupings	MAP
PRODUCT_INVENTORY	Keep track of remaining inventory of products	DIMENSION

II: Map

Map tables are almost always a bad idea. Whatever they can do, a dimension table does better. That being said let's look at our new map table.

PRODUCT_CATEGORY			
PRODUCT_ID	**PRODUCT_NAME**	**PRODUCT_GROUP_ID**	**PRODUCT_GROUP**
P001	Monopoly	PG1	Board Games
P002	Stratego	PG1	Board Games
P003	Grand Theft Auto 5	PG4	Video Games

We chose to do this as a Map table out of simplicity, and because we know toys almost never change categories.

Still, all of this info could go (and would be better suited) in our product dimension table.

But you can see the appeal here. It's just simple and easy to understand.

III: Dimension

PRODUCT_INVENTORY					
START_DATE	**END_DATE**	**PRODUCT_ID**	**AVAILABLE_QTY**	**SALE_PRICE**	**CLEARANCE**
20210301	99991231	P001	12	18.99	N
20210301	99991231	P002	15	9.99	Y
20210215	99991231	P003	100	49.99	N

Notice anything strange here?

This is a dimension table yet it contains values!

Can you guess why?

Well, if you walk into a store today and a day later…even a week later. Do you expect the price of a product to change? How often does Starbucks change the prices on their drinks? Not very often.

Because of this it's probably better suited to keep track of pricing in a long term way here in the dimension table.

Similar concept with quantity. While it's true that every time a sale occurs in the fact table, quantities are diminishing, each sale is only for a few products and our store has (presumably) thousands of products.

So if we sold one Monopoly (P001) just last week and none since, this quantity of 12 entry in the dimension table would have a START_DATE of immediately after that sale occurred to reflect the new quantity.

A good thing to keep in mind though, is that if price changes but quantity does not, that entry will still close and there will be a new entry in the table.

IV: Fact

SALES_HIST				
DATE	**TIMESTAMP**	**PRODUCT_ID**	**CUSTOMER_ID**	**QUANTITY**
20200101	20200101101525	P001	C001	2
20200101	20200101141820	P002	C002	1
20200101	20200101141842	P003	C003	6

So why are sales a fact table?

Think about it this way:

Every sale creates something NEW. The sale is a new entry. A new event occurred.

When we update a dimension or map table what are we doing? We are UPDATING a record.

You'll notice something new in this table that we don't have in our finance dataset, but this is pretty common in finance.

TIMESTAMP

What if a customer comes in TWICE in a day and purchase the same product? If we were only using date, there would be a duplicate row – it would look like bad data! But in reality there were actually two sales!

So we use TIMESTAMP which records the date and time so we will see this properly.

Here is how you read this TIMESTAMP:

20200101101525

Year = 2020

Month = 01

Day = 01

Hour = 10

Minute = 15

Second = 25

If we were creating a primary key for this table, something that uniquely and comprehensively describes each entry, it would probably be:

TIMESTAMP + PRODUCT_ID + CUSTOMER_ID

E.g. 20200101101525P001C001

IV: "[XYZ] Based Table"

Most tables have a GOAL. A purpose. A sales table lets us store the history of our sales. In our real dataset, we keep track of employees with employee info.

Similarly, some tables are constructed with single purposes in mind – to be able to do analysis on a certain cut of data.

If you wish to design a "COUNTRY-BASED FACT TABLE" here are the details you need to consider:

1. It must follow the frequency routine of the fact table. That is if it is a daily table, the countries must show up once a day. If it is a monthly table, then each country shows up once a month.
2. "Country Based" almost always means SUM everything under one country. That is if each country's trades can further be broken down by trader and client, sum that all up. Here is an example:

DATE	TRADER	CLIENT	COUNTRY	QUANTITY
20200102	T1	C01	France	300
20200102	T2	C03	France	6,600
20200102	T5	C04	Spain	8,300
20200102	T1	C04	USA	8,100
20200103	T7	C01	USA	7,500
20200103	T3	C03	USA	8,200
20200103	T4	C05	France	7,400
20200103	T1	C02	France	6,400
20200103	T1	C01	France	3,000

Above is a regular fact table you might see. Notice this table has a daily frequency. If we wanted to see this as a "Country Based" dataset, this is what we would find:

DATE	COUNTRY	QUANTITY
20200102	France	6,900
20200102	Spain	8,300
20200102	USA	8,100
20200103	France	16,800
20200103	USA	15,700

Notice there is no Spain on Jan 3rd. That's OK since there were no trades in Spain on that day.

3. It is *possible* that a "[XYZ] based table" might be averages, not sums. This would certainly be the case if the values we were looking at were RATES and not quantities, as summing up rates make no sense. However, for the purpose of this course, let's assume we are talking about SUM only for these type of questions.