**Privacy**

***Due to privacy policies I will NOT disclosure ANY real name, database, table and/or content. All the content is this post was changed to preserve privacy.***

**Scenario**

When I was hired and 12/april/2022, I accepted this very job due to the challenge that was offered to me, because I was looking for a job outside Brazil at the time. But… it was a huge challenge and I accepted it.

The company was approximately 80 employees, at the time I was hired, there was no structured IT Sector, just a bunch of guys in front of computers supporting the company.

In my first day it was given to me a lot of tasks and to each and every task that was given, I asked:

*Where is the data?*

Well… there was a lot of sources… ANY of the sources was a Database itself, it was tables in Google Drive, local excel tables and reports exported from a program that I will call it “ABC” from “X Company”.

This program was hosted in a Local Host PC located inside the my company and all of our computers were linked to it. And the most important informations for the company were in the program “ABC”. So, I had to manually export all the data from the “ABC” and make the dashboards and/or analysis.

Nonetheless, it was required from us (IT Sector) to automate everything. Then, we asked to “X Company” to move the Database of “ABC” to the cloud.

When it was finnaly moved to the cloud, then we had to connect a VPN to the cloud and download everything from the cloud to our own cloud (a Google Big Query). Well finally done, at the time, we started with approximately 450 tables and 5000 columns.

So, in order to start my job in this mess I asked my supervisor to ask “X Company” to send us the database diagram.

**Documeting an entire Database**

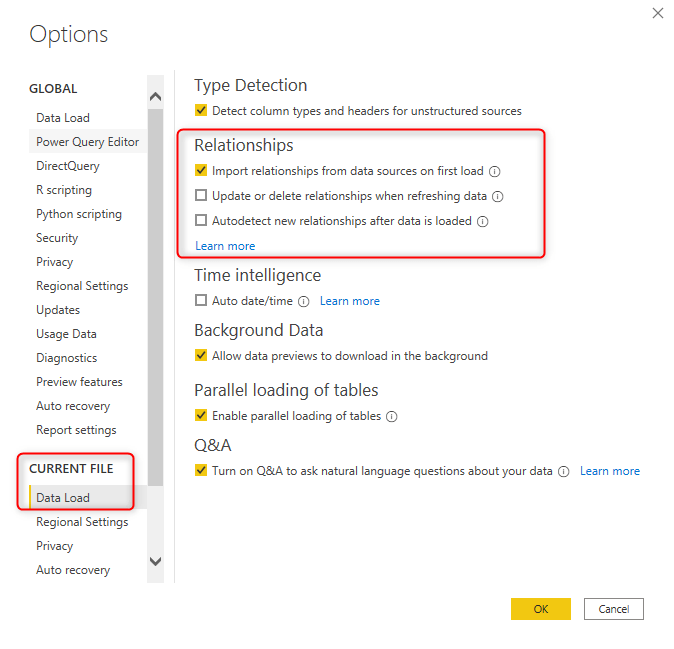
Their answer was: **We won’t, because it is our core business.**

Well, that’s where everything started. My supervisor asked me to document the database, because, without it, the most important projects in our company will be **halted.**(no pressure, at all)

*How do I begin?*

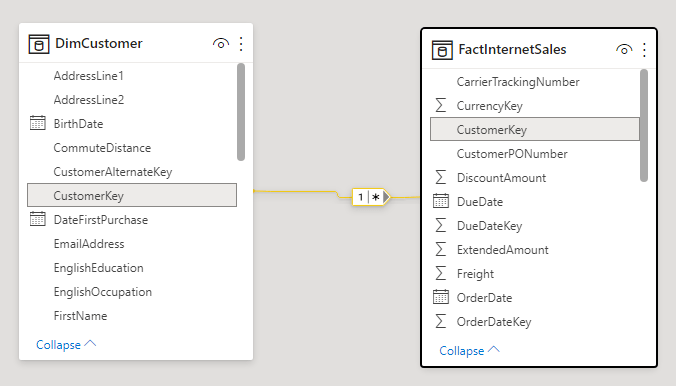
I have more than 400 tables, within the tables there are more than 5000 columns, with their FK ([foreign key](https://en.wikipedia.org/wiki/Foreign_key)).

I had an idea, Power BI can identify columns with the same name, if I check the box “Autodetect new relationships after data is loaded”.



Source: <https://radacad.com/power-bi-design-tip-disable-the-auto-detect-relationship>

With this field checked, the Power BI is able to do it:



Source: <https://radacad.com/power-bi-design-tip-disable-the-auto-detect-relationship>

Therefore I imagined… that’s it. Problem solved.

*Well… it does not.*

The nomenclature of the database is not standardized, but this idea gave me a good start. When I loaded the entire database in Power BI, I was able to recognize the there were some tables that were linked to more tables than another tables.

Aware of it, I decided to start the documentation separating [Fact Tables](https://en.wikipedia.org/wiki/Fact_table) from [Dimension Tables](https://en.wikipedia.org/wiki/Dimension_(data_warehouse)). Summerizing it, according to ChatGPT: *“The fact table typically contains foreign keys that connect it to dimension tables, which provide additional descriptive information about the facts.”*

Before going further, I’d to explain why I made this decision: Star Schema.

In a Star/Snowflake Schema the Fact Table is connected to the Dimension Tables, therefore, the tables that shows the highest number of connections in PBI must be the Fact Tables and the least ones, must be the Dimension Tables.



Source: <https://www.databricks.com/glossary/star-schema>

I remember now, how I reacted when I realized the size of the effort that would be required not only from me, but many people within the company.

**“A Mission Given is a Mission Accomplished”**

First of all, the tools:

1. Google Big Query (SQL) (hundreds of thousands of queries)
2. Power BI
3. Google Docs
4. Diagrams.net (which can be linked with Google Docs)

At that time, we knew (IT Sector, Managers from all levels) that every project depends on my “mission”, luckly their awareness was extremely helpful for me, because “we”, as an organization were entirely focused to achieve a single objective, which was a pretty unique moment in my life.

To achieve our objective, that is, to document the database, it was required a lot of people, many hours of work, countless reports and endless meetings of validation.

At that very moment, everything was ready to start the documention proccess.

First of all, I started looking for all the fact tables that I could identify, I didn’t correctly identified all of them at first, but I was able to recognize the 4 main tables, which I will designate as table\_a, table\_b, table\_c and table\_d.

Those tables represent the core business of our compay and brings all the information that was required for many project that were halted.

After the “ID confirmation” I was able to proceed and link them with their respective dimension tables.

But another problem showed up. There were more than 100 dimension tables linked to these 4 tables and to make things even worse, the nomenclature of most of the tables were compleletly non-standardized. For an example: table\_b.random\_variable does no match with a dimension table called “random\_variable”, just few tables have a standardized nomenclature as table\_b.variable\_abc matched with table variable\_abc.

At first, to check if the tables were correctly linked I tried to visualize if all the values within the dimension table were at the fact table, but many of the values were abandoned of deleted from the dimension table, which leaded me to a lot of errors, then I tried another approach. I started asking for reports from the program “ABC” and started to validate it.

There were a lot of other tables that was required video calls with different people with different background of different sectors within the company.

Documenting the 4 main tables of the database allowed us (the company) to move some projects foward, but the database wasn’t even close to be decently documented.

All this process took almost 2 months of work, with full time dedication (9 hours per day, 5 days a week).

**Finance tables**

With the operational area documented (not 100%, as I said), I was assign with a new task, and the most difficult one: document the finance tables. At first it was apparently easy, but as the time goes by it became a nightmare of complexity.

When this task was assigned to me, I was already aware which tables were the tables related to finances.

Then I repeated the same process I did before: reports, video calls, validation and on and on and on…

Great, but finances itself is a complicated area, so summarizing it, a year after: there are 2 main tables which I’ll call fin\_1 and fin\_2. The table fin\_1 and fin\_2, they are basically the same thing, almost the same columns, same dimension tables… almost everything and both of them are active. You could be asking youself: why? … Well, I don’t know. But that’s how things work. We managed to figure it out and today it is working perfectly.

**Summarizing**

To document the basics about finances and operational took me 3 months of work, but only in december/2022 I finally delivered what was promissed in march/may of 2022 due to other urgent demands that eventually showed up for me.

Today, The documentation itself follows the same structure that I created back then. I believe that approximetly 95% of the database is properly documented in a 50 pages google docs document, with tens of diagrams linking hundreds of dimension tables to its fact tables, explaining in text which are the columns, the foreign key, which information they bring and a lot of more “structural” information about the database.

The database changes almost every month, and updates are being made to the documentation as soon as any data, table and/or information is added to our database.

**Special thanks**

I would like to say thanks to each and every person (I believe that up to 50 people) that helped me throught this project that still going. This project is a remarkable milestone in my professional career.

**Igor Comune.**