**Introdução:**

Objetivo do projeto é ler 3 arquivos csv, fazer tratamento necessário e criar um relatório consumindo os dados inseridos.

**Dados origens:**

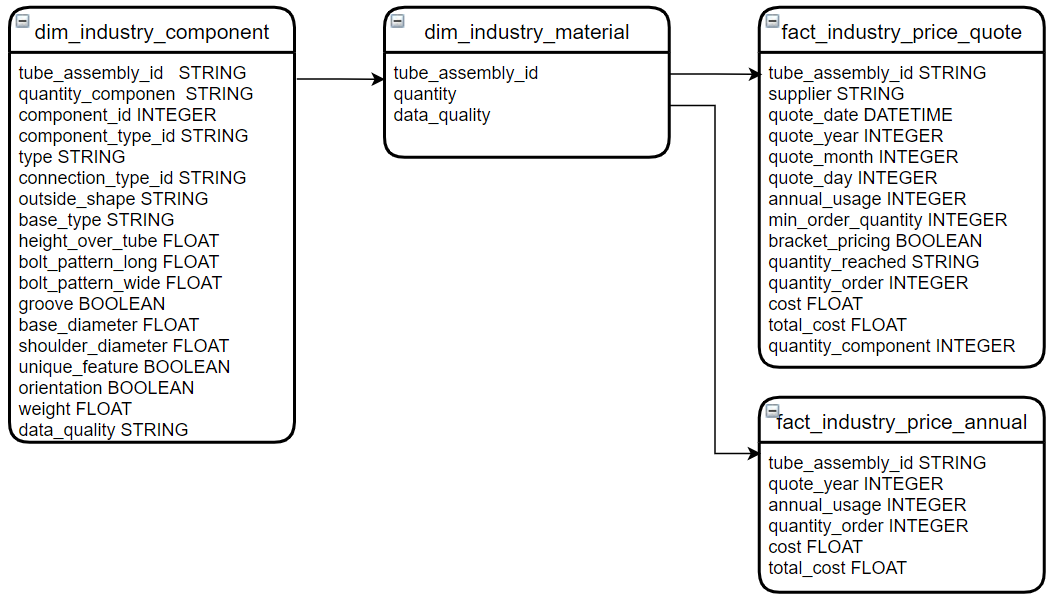
* price\_quote.csv
* bill\_of\_materials.csv
* comp\_boss.csv

**Análise:**

A solução adotada foi trabalhar o dado no conceito de data lake com 3 camadas:

|  |  |  |
| --- | --- | --- |
| **Raw Zone** | **Trusted Zone** | **Refined Zone** |
| A tecnologia usada foi armazenar os arquivos .csv com dados brutos no Google Clould Store. | A tecnologia usada foi ler os arquivos csv do Google Cloud Store via Data Prep. Realizando a padronização dos dados e aplicando unpivot.  Então os dados foram gravados em tabelas da camada trusted do Big Query. | Para a camada refined, foram criadas view materializadas utilizando o modelagem dimensonal snowflake, aplicando regras de agregação e preenchimento dos dados ausentes nas dimensões que estavam na fato.  O Data Studio acessa os dados desta camada para criação do dashboard. |

**Modelo Conceitual Refined Zone:**

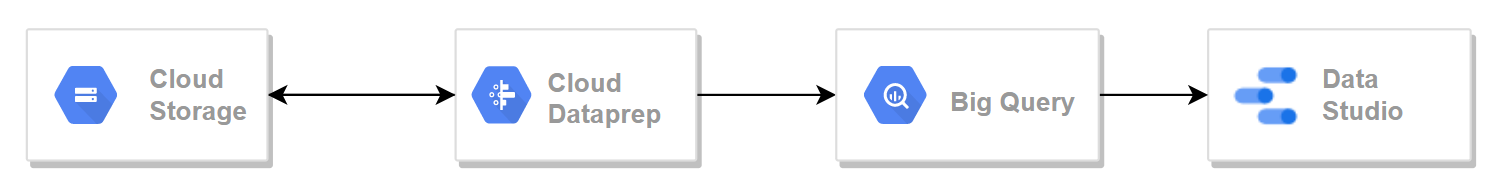


2. Diagrama ou Modelo dos dados (principalmente da trusted e refined) explicando para essa última os motivos de separar component\_id de material

3. Uma explicação sobre a qualidade dos dados e o que foi feito na refined a fim de mitigar e evidenciar o problema.

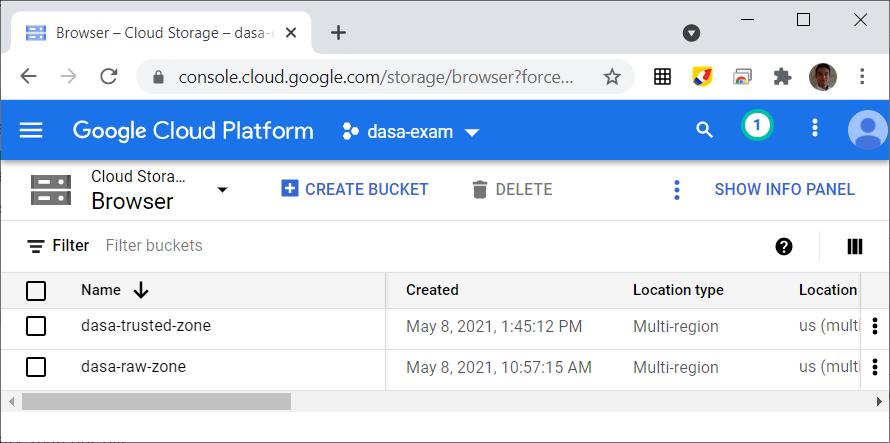
**Arquitetura da solução:**

**A arquitetura proposta contém 4 produtos do GCP:**



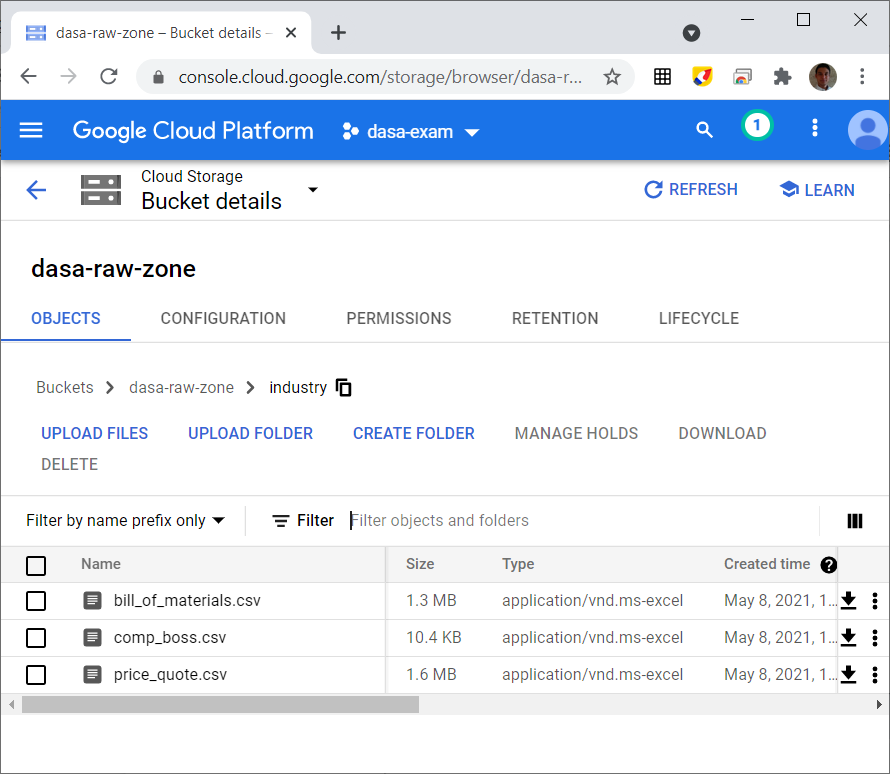
1. Arquivos origens são carregados no Cloud Storage no bucket raw zone
2. No Cloud Dataprep é feito ETL onde são aplicados os tratamentos necessários e os dados são gravados no bucket trusted zone
3. Processo é executado via 3 Flows do DataPrep, que são atualizados a cada 20 minutos.
4. No BigQuery foi criado o modelo dimensional e gravado no bucket refined zone
5. No Data Studio foram criados os relatórios lendo os dados da camada refined do Bigquery.

**1 - Clould Storage:**

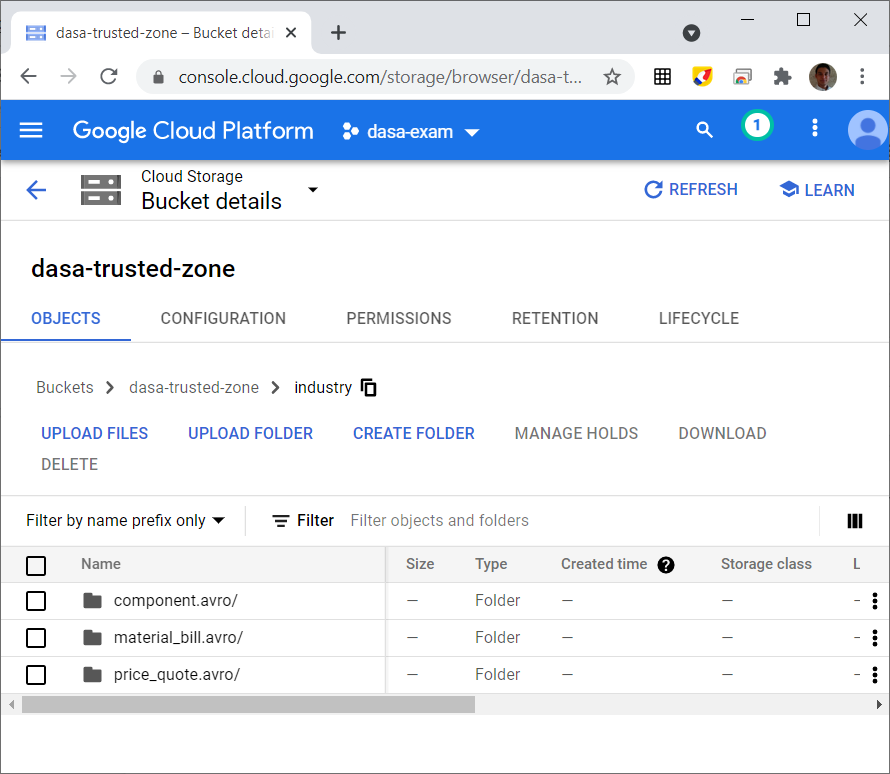
****

**Bucket:**

## dasa-raw-zone

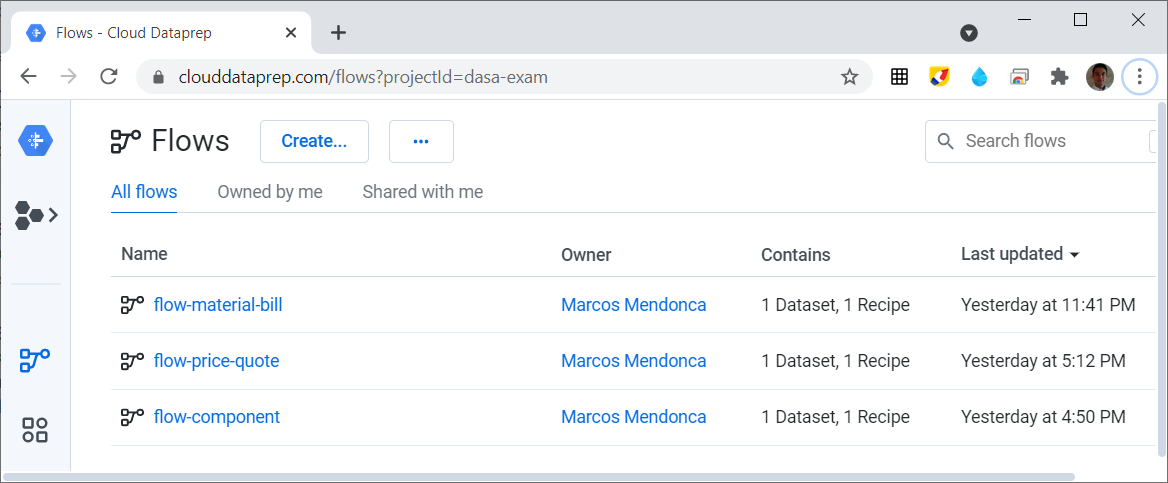


## dasa-trusted-zone

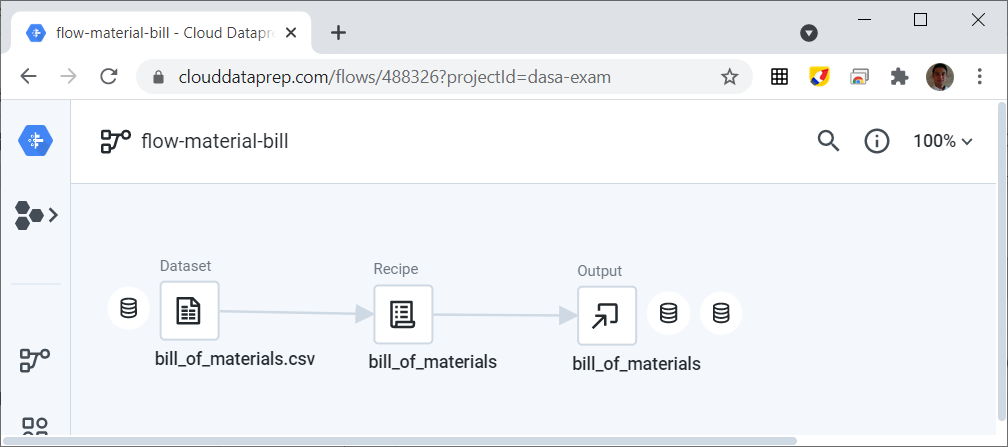


## 

**2 - Cloud Dataprep:**

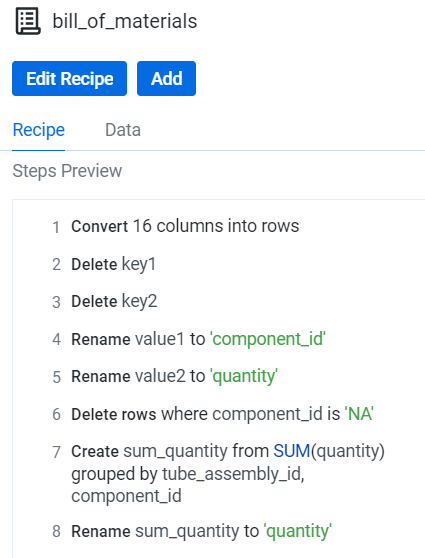


* **flow-material-bill**

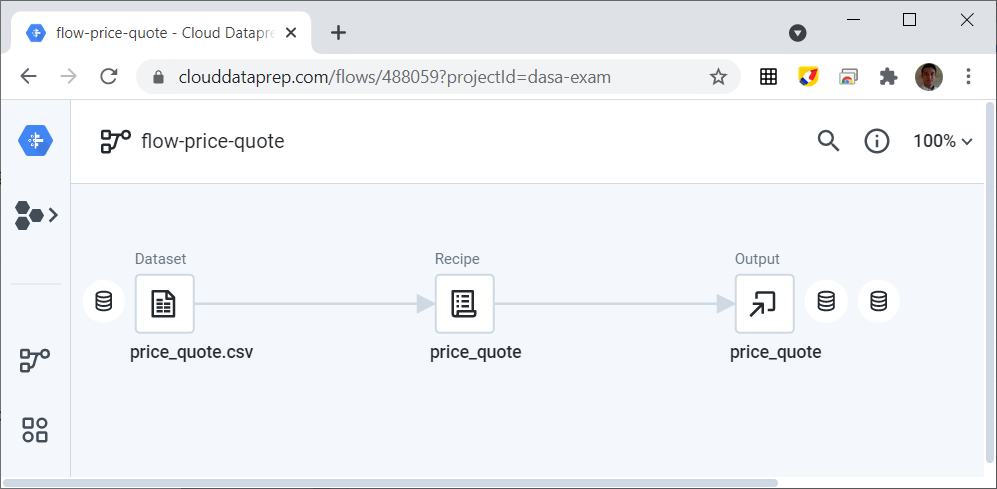


**Tratamentos:**

* Faz unpivot transformando as colunas de componentes e quantidades em linhas;
* Conversão dos data types;
* Limpa os valores N/A;
* Agrupamento por tube\_assembly\_id, component\_id e sumarização da quantidade (quantity), para eliminar duplicidade.

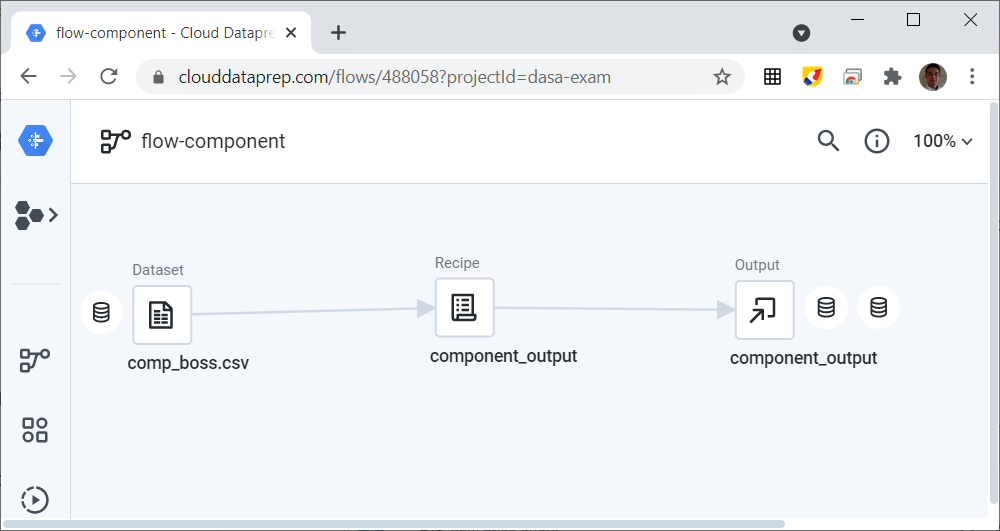


* **flow-price-quote**



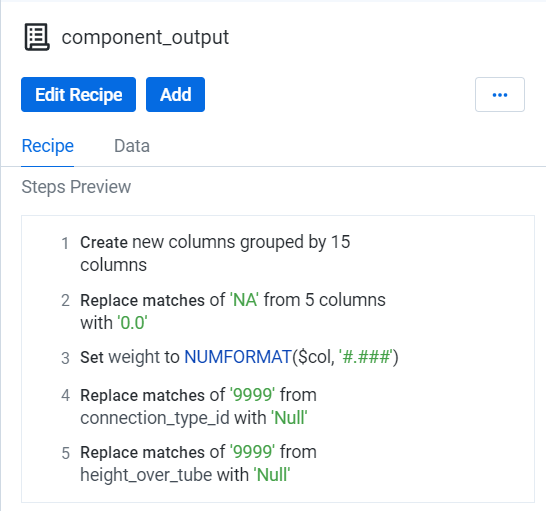
**Tratamentos:**

* Conversão de data types
* **flow-component**

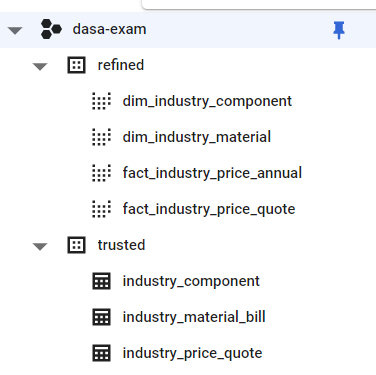
****

***Tratamentos:***

* Conversão de data types
* Remoção de possíveis registros duplicados
* Tratamento dos valores 9999 e NA para nulo

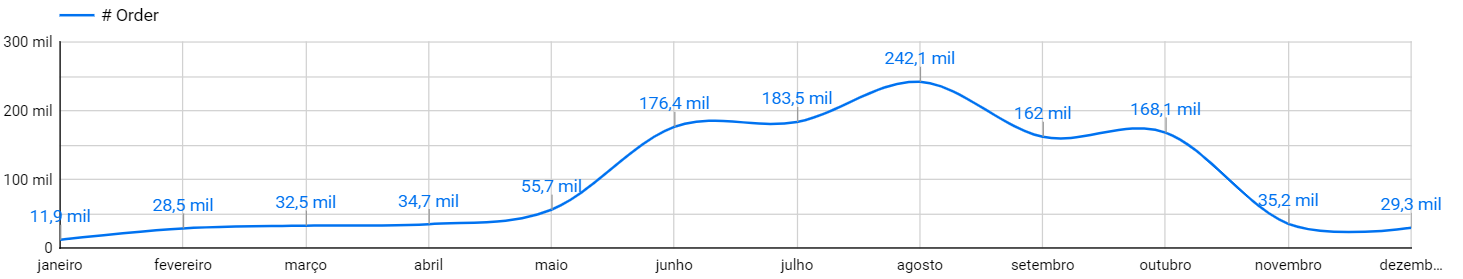


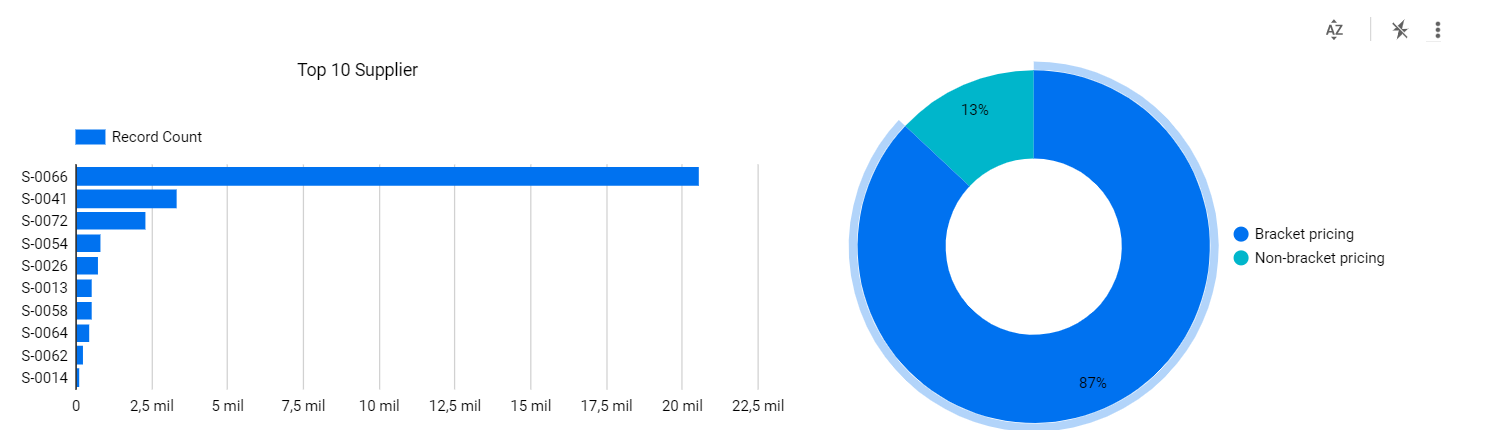
**3 - Bigquery**

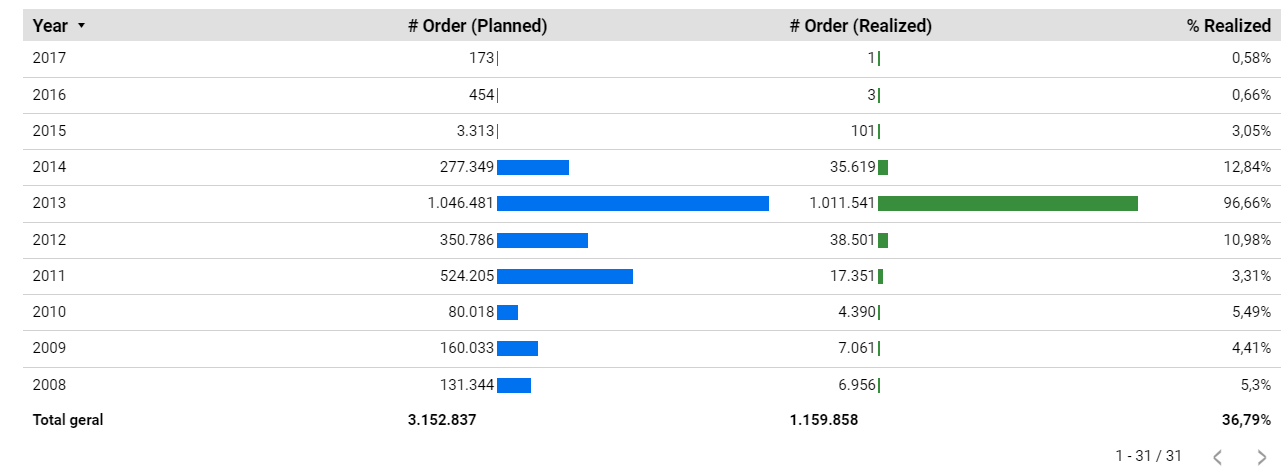
****

**4 - Data Studio:**

****

****

****



**Ao analisar o dashboard podemos concluir que:**

* O ano com maior número de ordens foi 2013.
* No consolidado o mês com maior número de ordens foi Maio.
* A forma de cotação Bracket pricing corresponde a 87% das cotações.
* Apenas 36% da meta foi atingida no consolidado dos anos.

**Possíveis melhorias:**

* Acrescentar o Google Cloud Composer para orquestração do fluxo do pipeline.
* Criar um processo de validação e aviso dos possíveis dados ausentes dos cadastros aos data owner dos dados.