MVP – Minimum Viable Product

Pipeline Engenharia de Dados

Índice

Sumário

[**Controle de Versão** 3](#_Toc171635573)

[**1.** **Objetivo** 4](#_Toc171635574)

[**2. Plataforma** 4](#_Toc171635581)

[**3. Detalhamento** 4](#_Toc171635582)

[**3.1 Busca Pelos Dados** 4](#_Toc171635583)

[**3.2 Coleta** 4](#_Toc171635584)

[**3.3 Modelagem** 6](#_Toc171635585)

[**Catálogo dos dados:** 6](#_Toc171635586)

[**linhagem dos dados:** 6](#_Toc171635587)

[**4. Carga** 6](#_Toc171635588)

[**5. Analise** 7](#_Toc171635589)

[**6. Arquitetura** 8](#_Toc171635599)

[**7. Fluxo Pipeline** 9](#_Toc171635600)

[**8. AutoAvaliação** 12](#_Toc171635601)

# Controle de Versão

Nessa seção, serão descritas todas as atualizações realizadas no documento, bem como a data e o responsável por elas.

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Versão | Responsável | Atualização |
| 12/06/2024 | 1.0 | Leonardo Silva | Layout Documento |
| 13/06/2024 | 1.1 | Leonardo Silva | Estudo sobre o tema. |
| 16/06/2024 | 1.2 | Leonardo Silva | índice, descrição inicial. |
| 22/06/2024 | 1.3 | Leonardo Silva | Ferramentas Web GCP |
| 01/07/2024 | 1.4 | Leonardo Silva | Config Metabase |
| 04/07/2024 | 1.5 | Leonardo Silva | Catalogo de dados |
| 06/07/2024 | 1.5 | Leonardo Silva | Ferramentas Web BigQuery / Datalfow |
| 07/07/2024 | 1,6 | Leonardo Silva | Imagens e estrutura de entrega |
| 08/07/2024 | 1.7 | Leonardo Silva | execuçao fluxo pipeline e imagens |
| 10/07/2024 | 1.8 | Leonardo Silva | Revisão e AutoAvaliação |

# Objetivo

O projeto tem o objetivo desenvolver uma pipeline de dados utilizando tecnologias na nuvem. Essa pipeline deve necessariamente envolver a busca, coleta, modelagem, carga e análise dos dados, bem como responder 5 perguntas relacionadas a emissão de CNH’s (Carteira Nacional de Habilitação), emitidas em 4 estados brasileiros, com base em uma volumetria/coleta de dados por período.   
  
Perguntas:

1. Total Emissão por UF.
2. Porcentagem % Emissão por UF.
3. Porcentagem % de emissão por TIPO CNH. (**P**ermissionário x **C**ondutor).

4- Media tipo categoria por idade.

5- Media idade primeira habilitação.

# 2. Plataforma

Direcionei os esforços de apoio na Plataforma **Dataproc**. O Dataproc é um serviço de nuvem rápido, fácil de usar e totalmente gerenciado para executar clusters do Apache Spark e Apache Hadoop com mais simplicidade. O Databricks Community Edition sugerido, possui uma particularidade na versão não muito amigável que tende a ficar recriando a todo momento a pipeline.

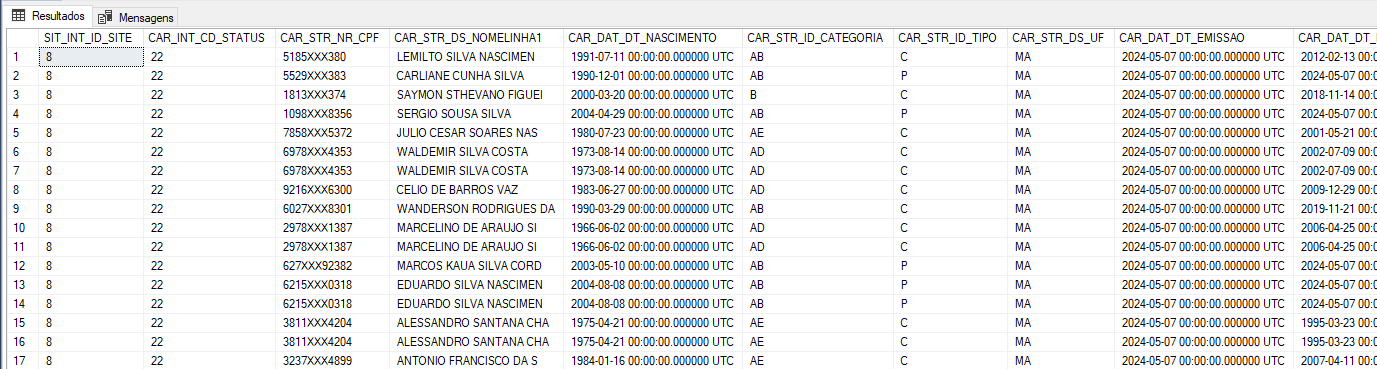
# 3. Detalhamento

## 3.1 Busca Pelos Dados

Iniciei um garimpo na busca por dataset’s gratuitos disponíveis na web, identifiquei alguns bons e interessantes, cheguei a iniciar uma análise dos benefícios por localidade (https://portaldatransparencia.gov.br/beneficios), porem me senti mais confortável em usar uma pequena massa de dados relacionado a emissão de CNH de 4 estados brasileiros.  
 ***\*\* Dados privados tratados devido LGPD.***

## 3.2 Coleta

A coleta inicial dos dados foi feita em uma base de dados SqlServer, composta pela consulta abaixo, trazendo dados de forma original e posteriormente exportados para um arquivo **.csv**.

SELECT \* FROM emissao\_cnh c where c.STATUS = 22; Clique >> [Dados\_Inicial](dados_inicial%201.JPG)  
 **Após geração do .cvs foi tratado alguns dados como CPF, registro e número CNH.**

## 3.3 Modelagem

Ao contrário do Modelo Estrela, será adotado a Flat Table, que consiste em uma tabela desnormalizada para o atingimento de objetivos específicos em ambientes de BI.

Sabemos que o Modelo Estrela é o mais amplamente utilizado, no entanto o modelo adequado depende de uma análise que envolve os pilares de tecnologia, segurança da informação e estratégia de negócios e objetivo.

Clique >> [Flat Model Table](modelagem%20Flat.JPG)

## Catálogo dos dados:

Catálogo de dados é um inventário organizado de ativos de dados na organização. Ele usa metadados para ajudar as organizações a gerenciarem seus dados. Também ajuda os profissionais de dados a coletar, organizar, acessar e enriquecer metadados para oferecer suporte à descoberta e governança de dados.

Clique >> [Catalogo de Dados](Catalogo%20de%20Dados%20Image.JPG)

## linhagem dos dados:

A linhagem de dados é o processo de acompanhamento do fluxo de dados durante um período de tempo, fornecendo uma visão clara de onde os dados se originaram, como mudaram e do destino final dentro do pipeline de dados. Tentei utilizar o Dataplex que inclui uma funcionalidade muito útil para configurar e executar verificações de qualidade de dados em recursos, como tabelas do BigQuery e arquivos do Cloud Storage. Porem tive dificuldades e fiz um rascunho.

Clique >> [Linhagem de Dados](Linhagem%20de%20dados%20Image.JPG)

# 4. Carga

A etapa de carga do dataset para Cloud Storage / Bucket, foi feita manualmente e será utilizada e atualizada pela pipeline desenvolvida utilizando Apache Spark.

Clique >> [Carga de Dados](Carga%20de%20Dados%20Bucket%20Image.JPG)

Clique >> [Pipeline.pdf](pipeline.pdf)

# 5. Analise

## Qualidade dos dados:

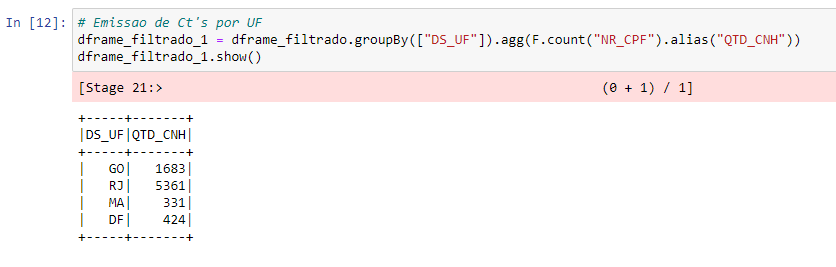
Os fatores qualitativos podem variar de caso para caso, mas de forma geral um dado de qualidade apresenta **completude**, **conformidade**, **precisão**, **consistência** e **integridade** (possui fontes e processos confiáveis e rastreáveis).

No conjunto inicial, os tipos de dados já apresentam padrões aceitáveis, mas para as respostas iniciais, devido uma particularidade, precisamos os campos **data** para **timestamp**, **string** para **integer**.

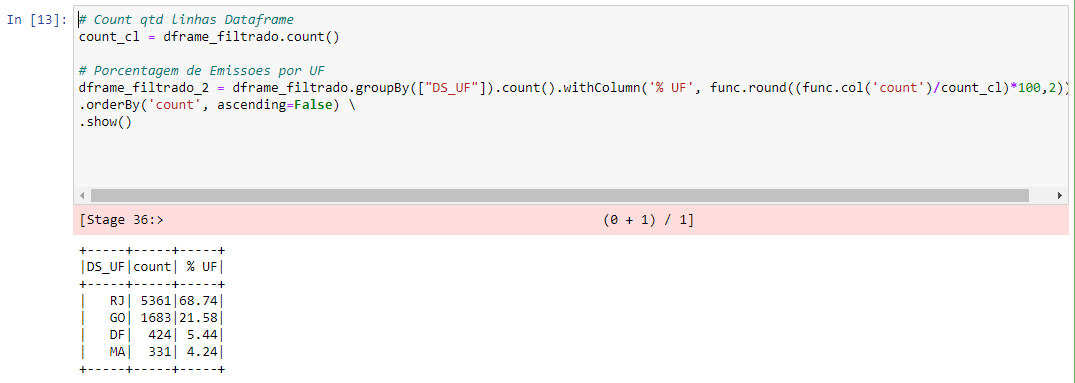
Alteramos o datatype pela pipeline de algumas colunas do DataSet para melhor representar as respostas iniciais.

Clique >> [Imagem Qualidade Dados](qualidade_dados.JPG)

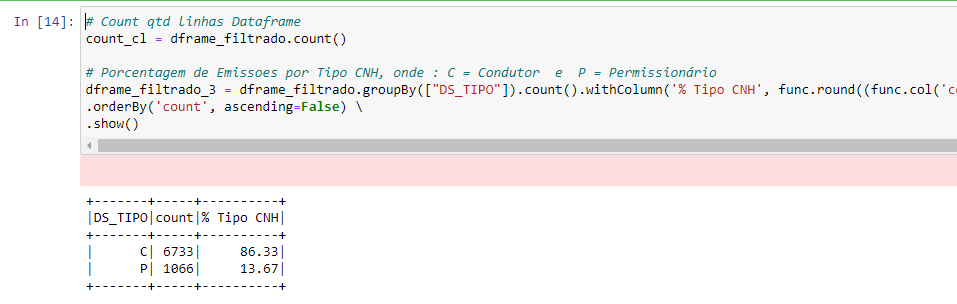
## Solução do Problema:

Respostas   
1- Total Emissão por UF

2- Porcentagem % Emissão por UF.



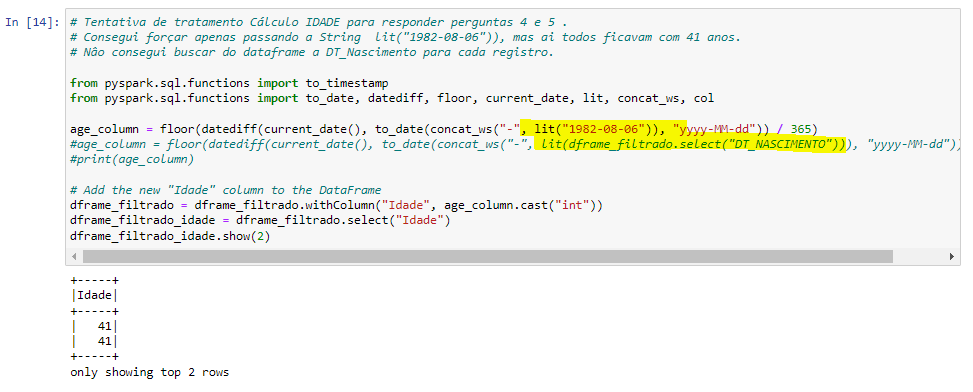
3- Porcentagem % de emissão por TIPO CNH. (**P**ermissionário x **C**ondutor).



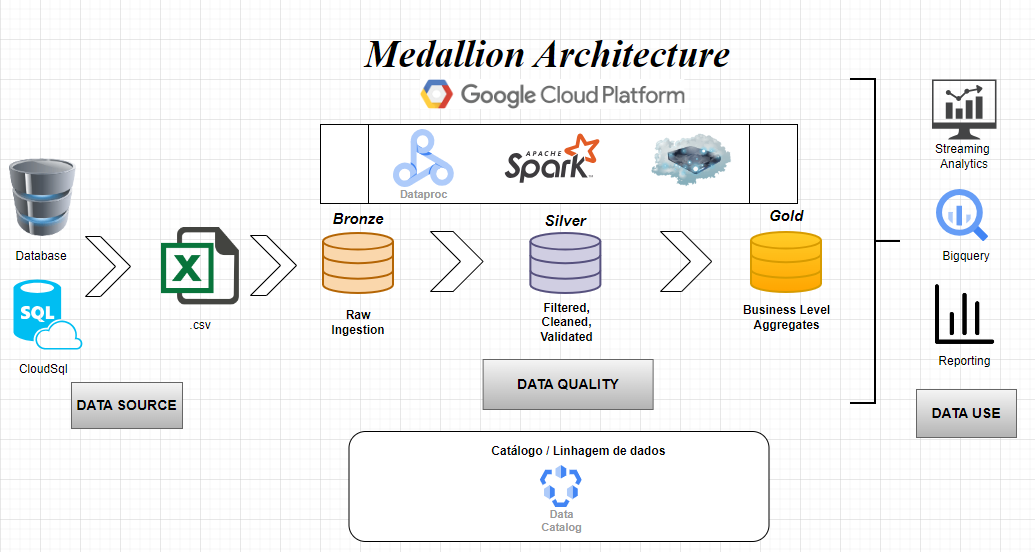
4- Media tipo categoria por idade.

> > Não respondida, pois na Pipeline não consegui tratar literal Idade.

5- Media idade primeira habilitação.

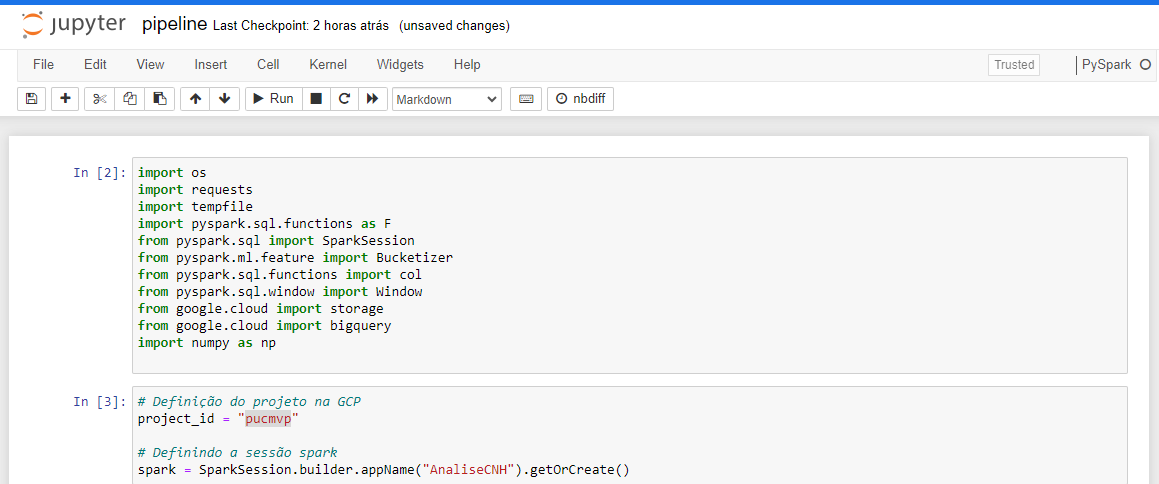
>> Não respondida, pois na Pipeline não consegui tratar literal Idade.  
  


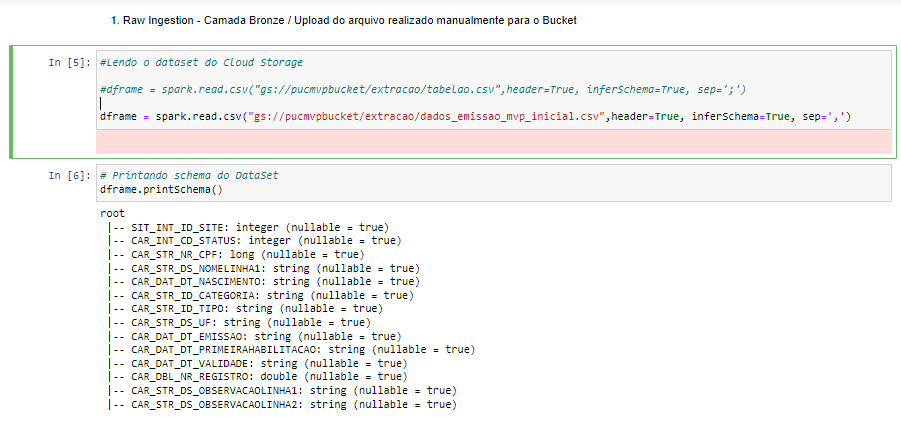
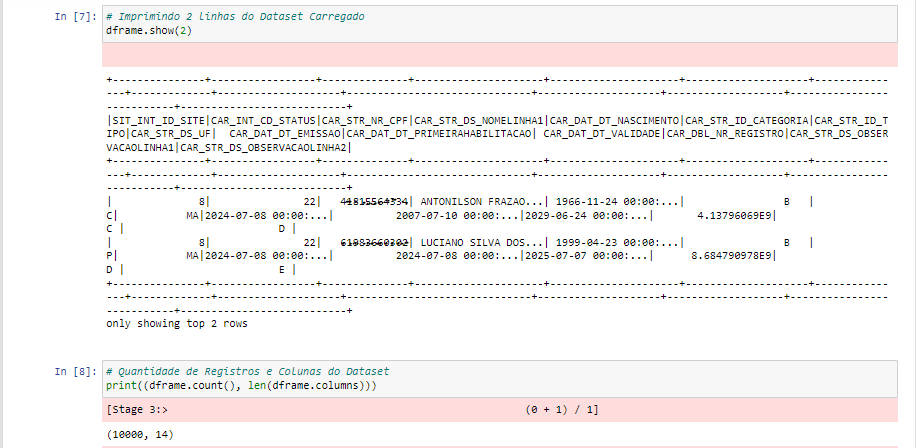
# 6. Arquitetura

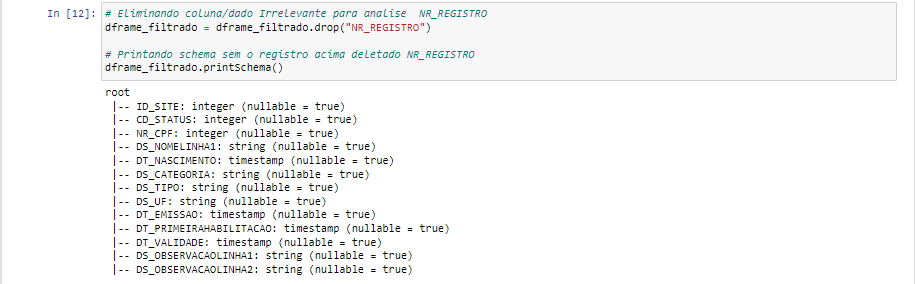
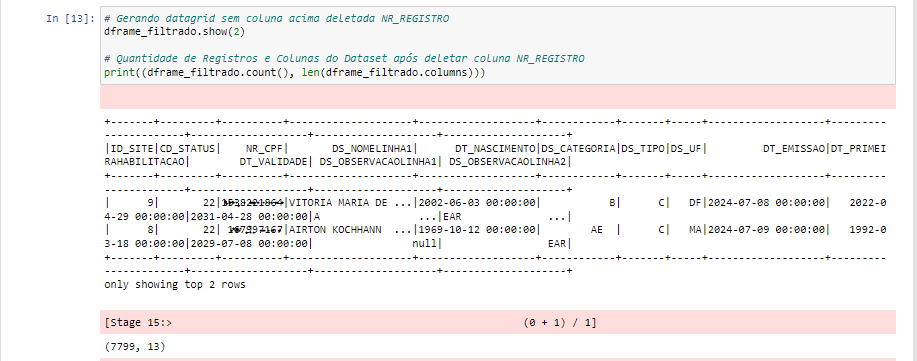
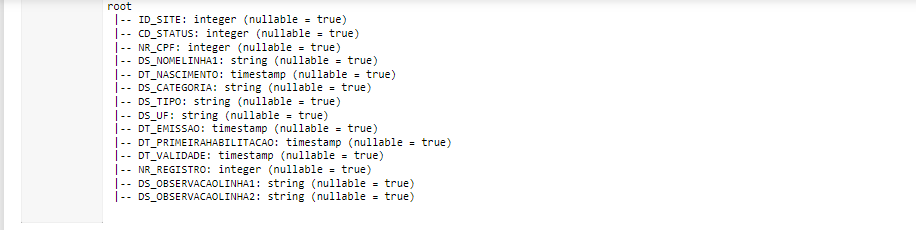
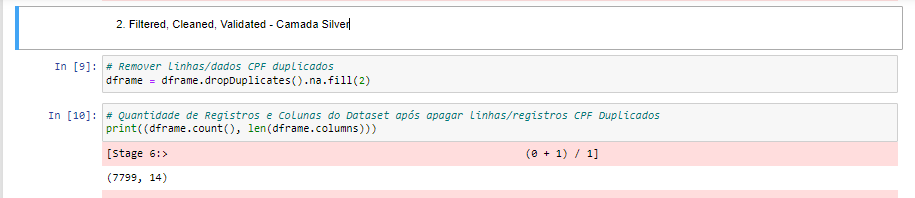


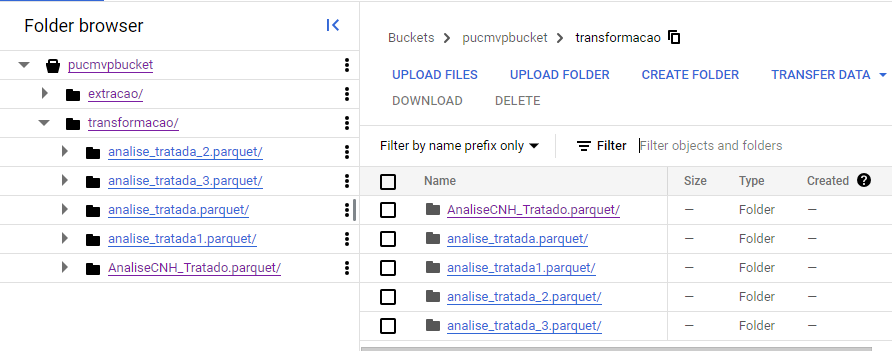
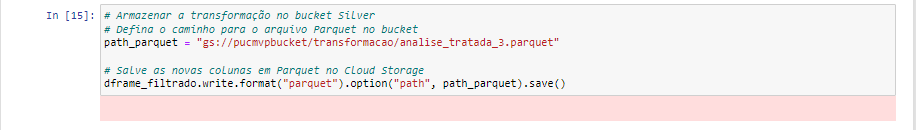
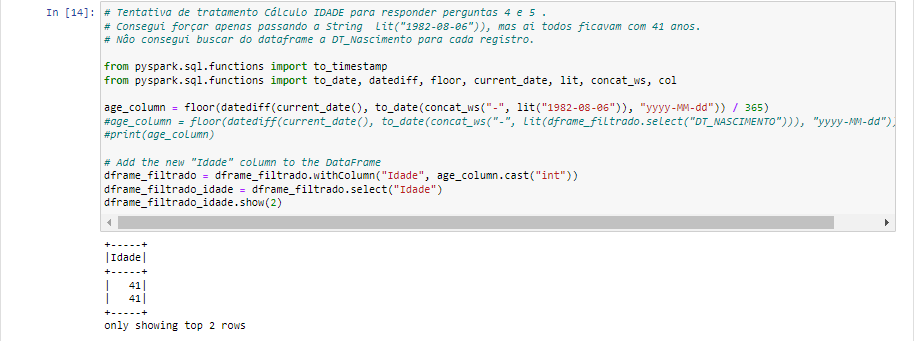
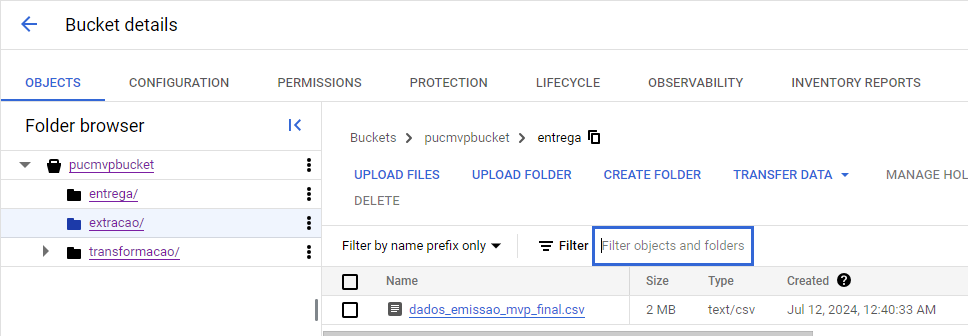
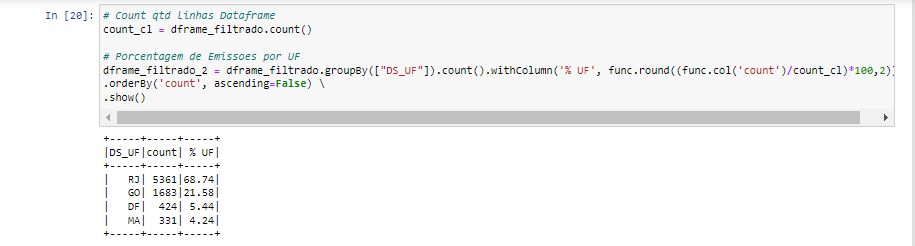
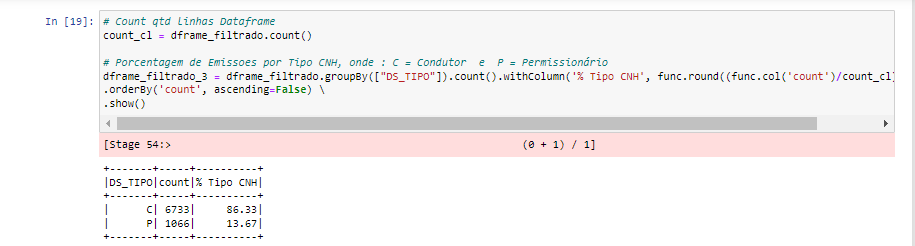
A arquitetura do medalhão exemplifica uma série de camadas de dados que denotam a qualidade dos dados armazenados. Recomendasse adotar uma abordagem multicamadas para criar uma única fonte de verdade para produtos de dados. Essa arquitetura garante atomicidade, consistência, isolamento e durabilidade à medida que os dados passam por várias camadas de validações e transformações antes de serem armazenados em uma disposição otimizada para análise eficiente. Os termos [bronze](https://docs.gcp.databricks.com/pt/lakehouse/medallion.html#bronze) (bruto), [prata](https://docs.gcp.databricks.com/pt/lakehouse/medallion.html#silver) (validado) e [ouro](https://docs.gcp.databricks.com/pt/lakehouse/medallion.html#gold) (enriquecido) descrevem a qualidade dos dados em cada uma dessas camadas.

# 7. Fluxo Pipeline

***Importação e Definição Project na GCP***

***Camada Bronze:***   
  


***Camada Prata:***

 ***Camada Ouro:***

# 8. Auto Avaliação

A ideia do projeto como um todo, é demonstrar uma abordagem bem estruturada e abrangente para o desenvolvimento de um pipeline tratamento/análise de dados de emissão de CNH. O projeto utiliza efetivamente tecnologias de nuvem, abordando aspectos de qualidade de dados e descrevendo um modelo claro para execução. As explicações e justificativas detalhadas para cada etapa indicam uma compreensão completa dos requisitos e desafios do projeto.

Inicialmente, fiz uma busca nos sites sugeridos para identificar dataset’s gratuitos, porem fiquei mais confortável em manipular e trabalhar com dados do meu dia-a-dia (emissão de carteiras de habilitação), pois as perguntas já eram bem definidas em minha cabeça.

Posteriormente, busquei diversos vídeos e tutoriais que descreviam e exemplificavam o que é e como criar um modelo de datawarehouse / datalake utilizando a linguagem Python e ferramentas web que facilitariam o desenvolvimento de uma pipeline.

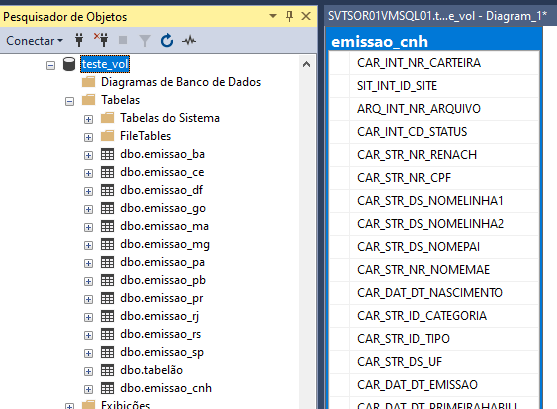
Cheguei nas ferramentas (Python, Dataproc, Dataplex, Nifi, BigQuery, Cloud Storage, Metabase, Datalog, Dataflow, etc .), e a sugerida pelos professores [ DataBricks ].

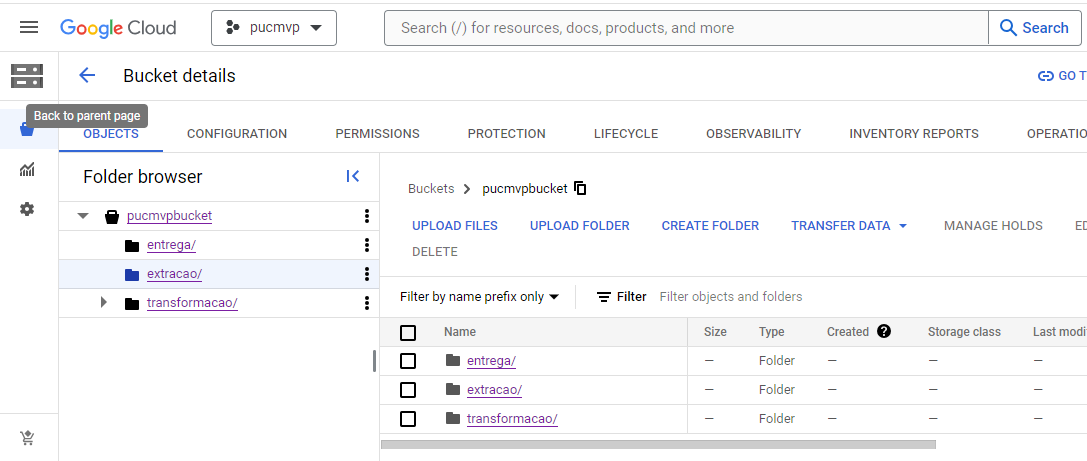
Tive bastante dificuldade em colocar na prática o fluxo para cada etapa: arquitetura, modelo e de análise dos dados, escolha do melhor dado, tratamento do dado em especifico e com isso minhas perguntas ficaram mais difíceis de responder.

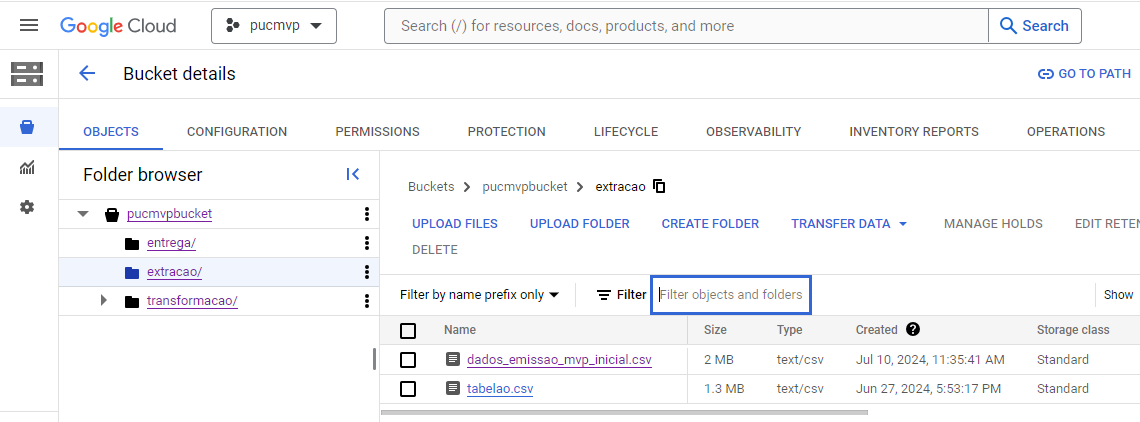
Optei pela Flat Table na criação do modelo, após muita dificuldade em criar as tabelas fato e dimensão. Nesse momento a ideia começou a ficar mais clara para o desenvolvimento (os encontros e o discord, tambem ajudaram muito a clarear as ideias).

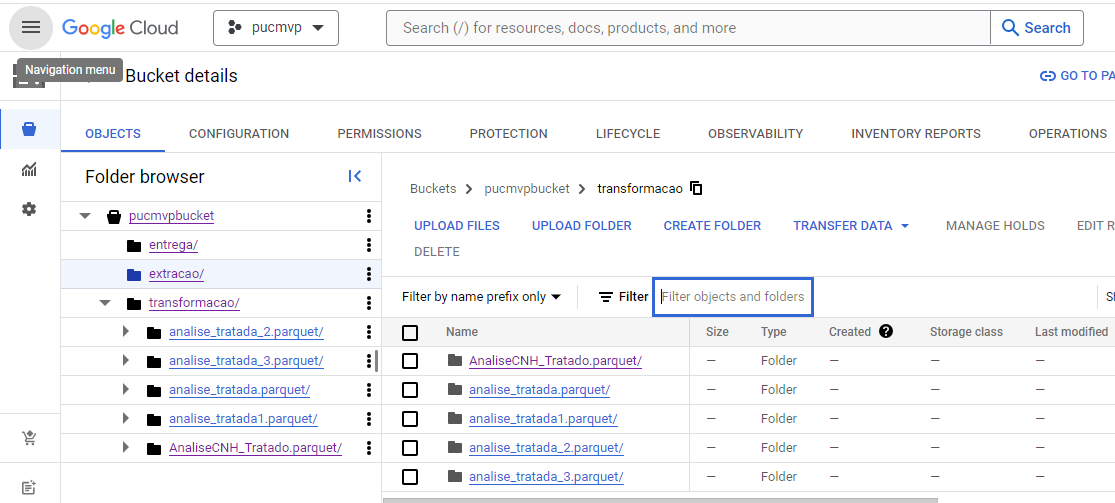
No contexto geral, fiquei muito satisfeito com as pesquisas e a capacidade de colocar em prática o aprendizado. Utilizei:  
   
**Cloud Storage** - Para criar o bucket e as estruturas de arquivos extração, transformação e entrega.  
**DataProc** - Para cria o cluster e usar a web interface do júpiter para desenvolver a pipeline.  
**BigQuery** - Para criar a estruturas de tabelas, catálogo de dados, carga e consultas dos dados.  
**DataPlex** - Para tentar criar a linhagem dos dados.   
**Metabase** - Para ler o dataset no BigQuery e gerar dashboards das respostas.

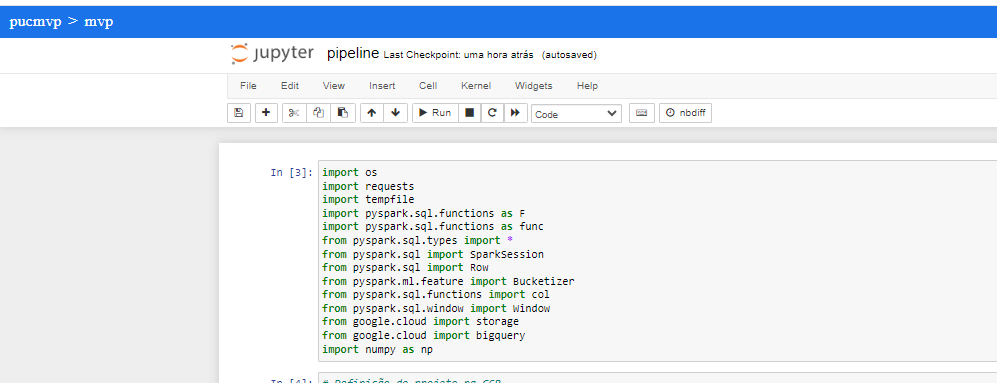
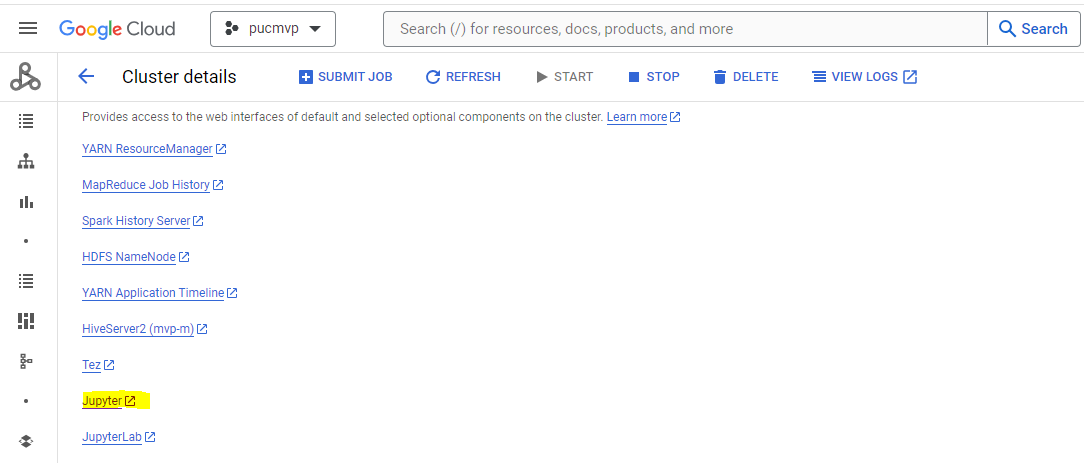
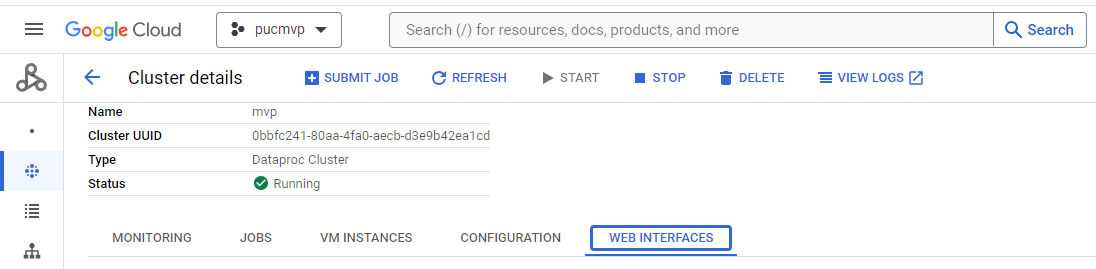
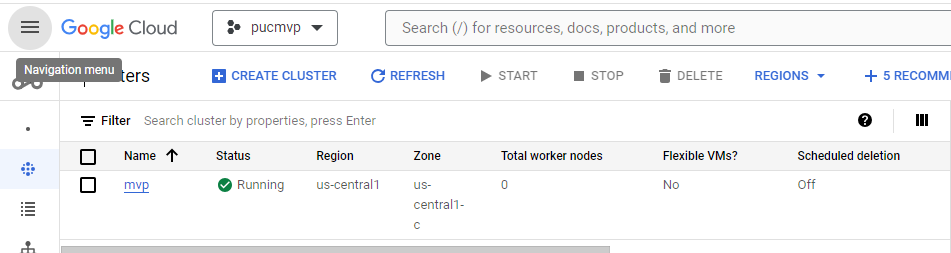
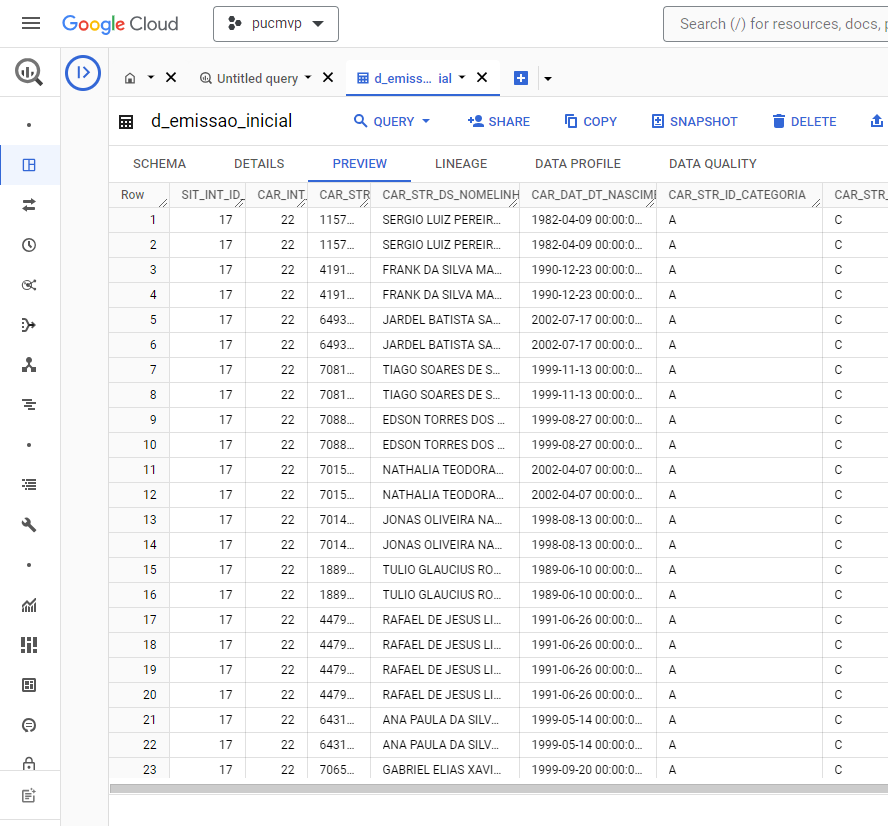
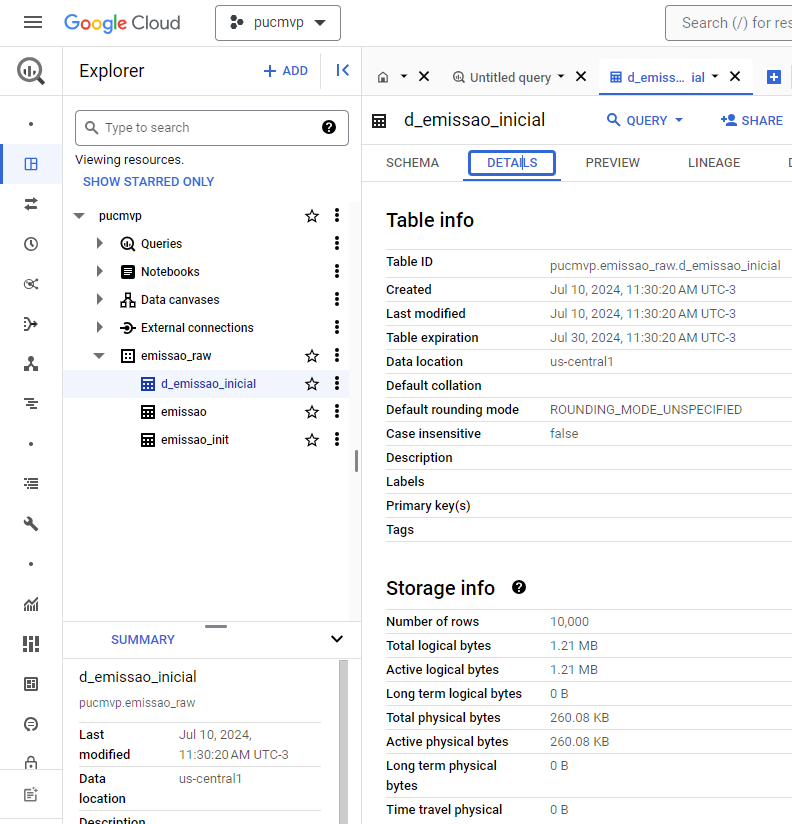
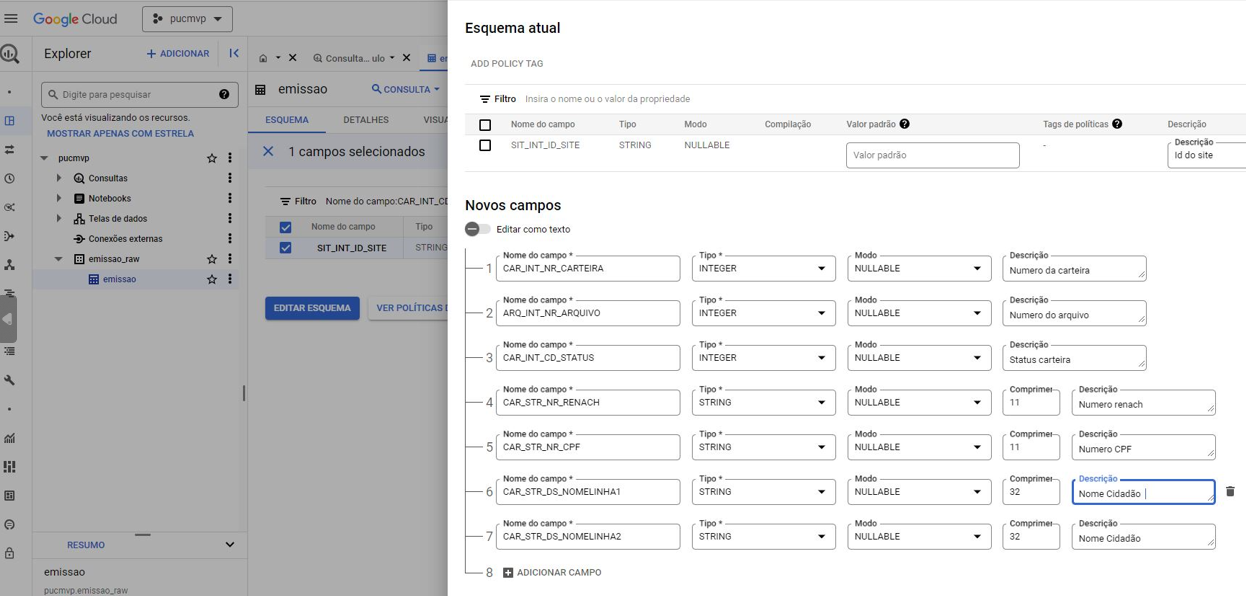
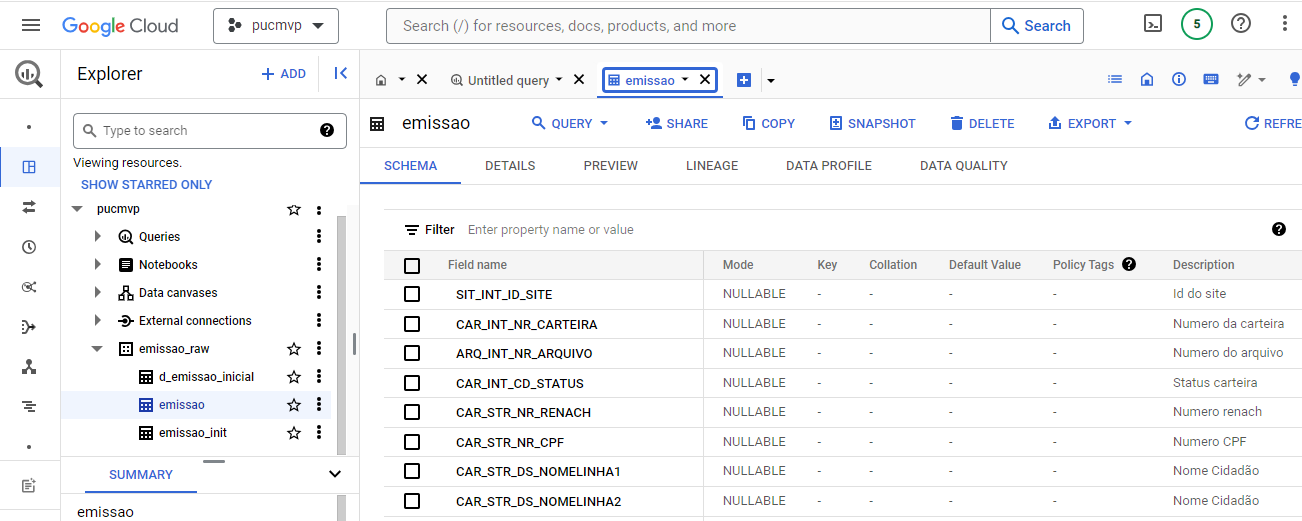
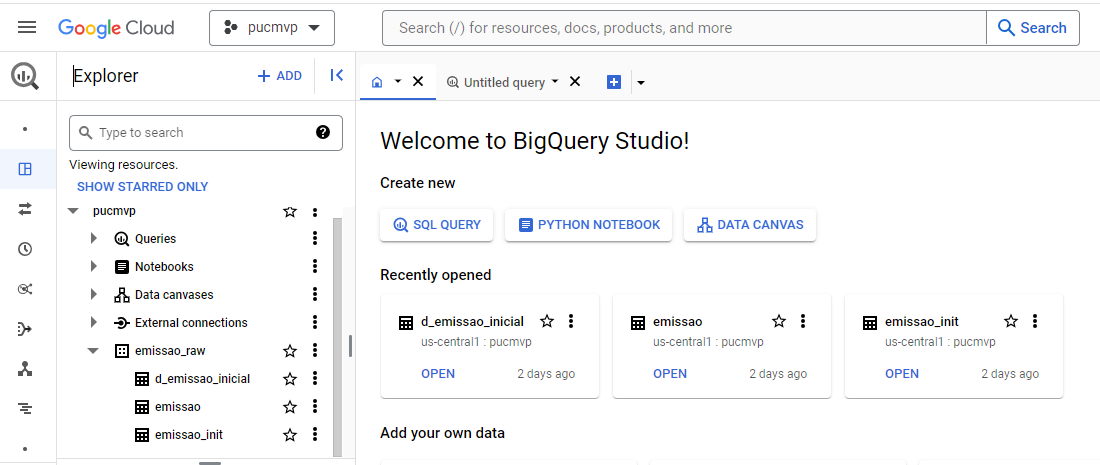
Algumas Imagens / Evidências das Ferramentas Utilizadas para desenvolvimento do MVP

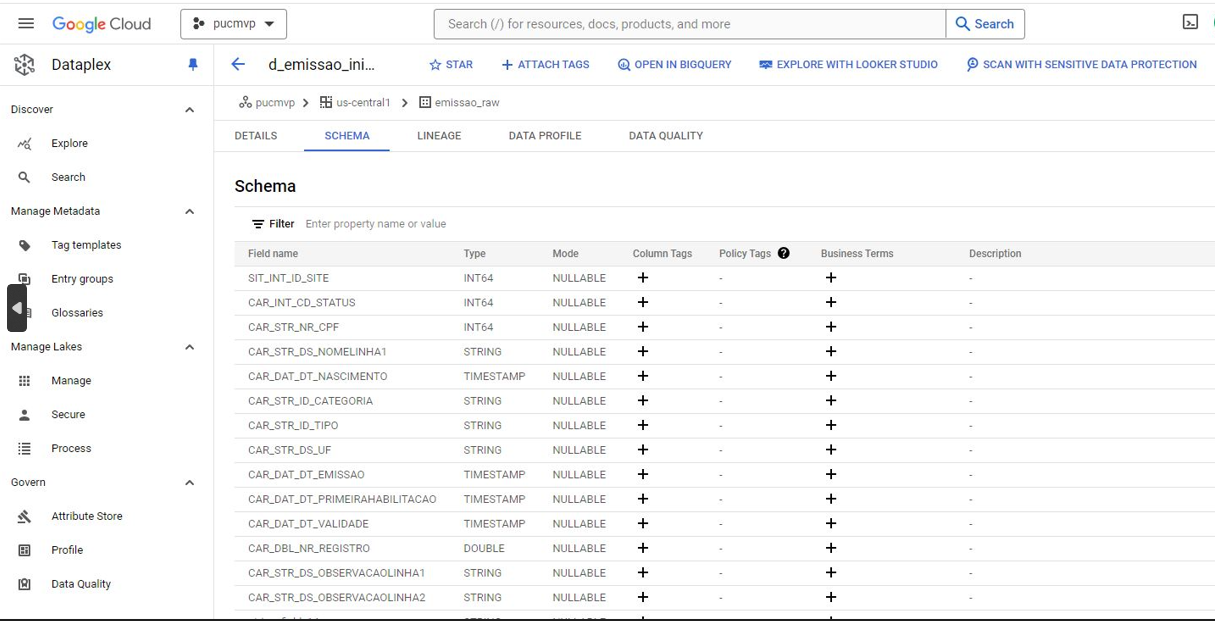
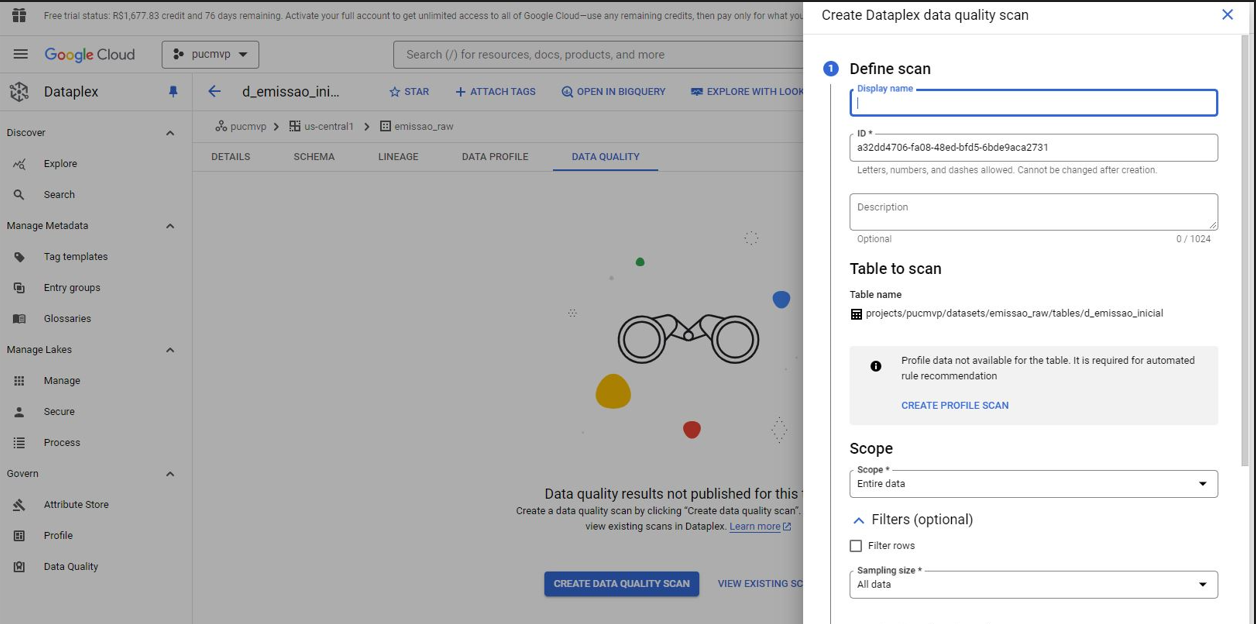
**SqlServer Management Studio**

**Cloud Storage Image**

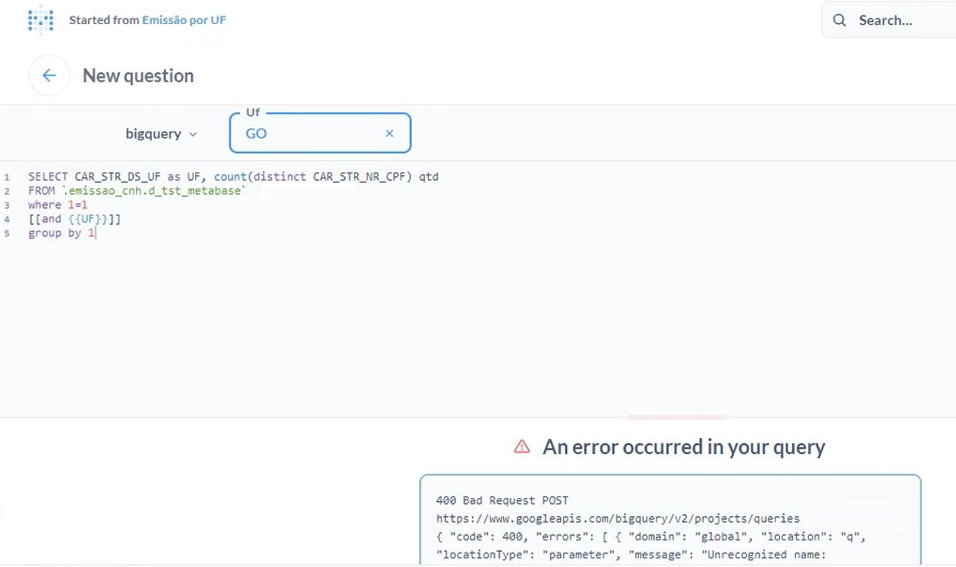
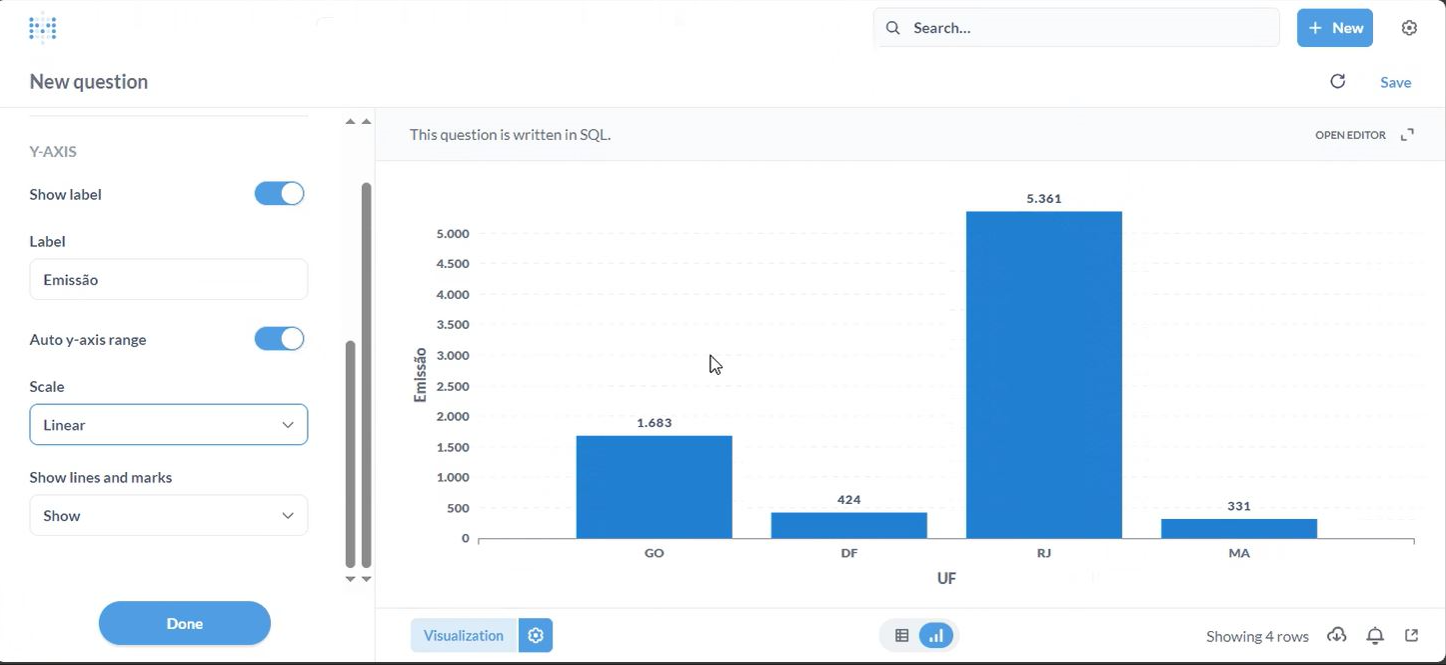
****



**Dataproc Cluster Image** **BigQuery Image**

**DataPlex**

**Metabase**

****

Respostas   
1- Total Emissão por UF