

MVP

Sprint: Engenharia de Dados

Base de Dados da Agência Nacional do Petróleo

Aluno: Leonardo da Silva Alexandre

Rio de Janeiro, 2023

Objetivo – Analisar o Valor dos Dados Sísmicos:

O mercado de óleo e gás brasileiro vive em constante movimentação e pesquisa para o desenvolvimento e descoberta de novos campos de petróleo. As pesquisas realizadas na busca de petróleo e gás geram um volume considerável de dados, onde parte destes dados são de origem geofísica. A partir de métodos indiretos a geofísica trabalha intensamente na busca de potenciais reservatórios em subsuperfície e no monitoramento dos reservatórios existentes.

Vários métodos geofísicos fazem parte desta cadeia de pesquisa, métodos potenciais, métodos eletromagnéticos, métodos sísmicos etc. O maior investimento ocorre na utilização do método sísmico, milhões de dólares são gastos em aquisições sísmicas que ocorrem por todo território nacional, tanto nas bacias terrestres quanto nas bacias marítimas.

Por lei, todo dado gerado pela indústria de óleo e gás deve ser entregue à Agência Nacional de Petróleo, Gás e Biocombustíveis (ANP), o que faz da ANP possuir um repositório de grande proporção para a indústria além de atuar como órgão regulador para o governo brasileiro.

Na organização desses dados existe toda uma legislação específica, uma parte delas discorre sobre a publicidade dos dados geofísicos, o que faz com que após alguns anos, período que varia de acordo com o tipo de dado, o dado geofísico se torna público e pode ser vendido pelo governo através da ANP.

Deixando de lado a questão legislativa sobre o tempo de confidencialidade desses dados, com foco apenas na parte matemática e financeira da questão, seguem as perguntas que guiam o objetivo deste MVP.

- 1. A ANP negocia esses dados a um valor coerente? Dentro do valor gasto para adquirir dados sísmicos, o valor cobrado pela agência é de fato compatível com o que foi gasto para criar esses dados?**
- 2. Uma empresa chegando ao Brasil hoje, é mais econômico comprar os dados da ANP ou fazer seu próprio levantamento geofísico?**
- 3. É possível para uma empresa igualar seu banco de dados com o da ANP?**
- 4. Quanto uma empresa gastará ao longo dos próximos anos comprando os dados que a ANP para manter seu banco atualizado?**

Detalhamento

1. Busca pelos Dados

Os dados foram coletados de partes diferentes dos sites da ANP. A planilha final foi construída anteriormente devido ao foco em pesquisas de cunho pessoal e profissional.

Os Dados foram extraídos a partir do shapefile dos programas geofísicos da ANP e da tabela de programas geofísicos.

- <https://www.gov.br/anp/pt-br/assuntos/exploracao-e-producao-de-oleo-e-gas/dados-tecnicos/acervo-de-dados>



Figura 1 - Site ANP onde pode ser encontrado o shapefile dos dados.

2. Coleta

Dados baixados para a máquina local, plotados utilizando o software Qgis e extraída a planilha de atributo do arquivo shapefile. Com a planilha fato criada com as informações oriundas do arquivo shapefile, os dados foram carregados manualmente no bucket “mvp-leoalex” no Cloud Storage da Google Cloud.

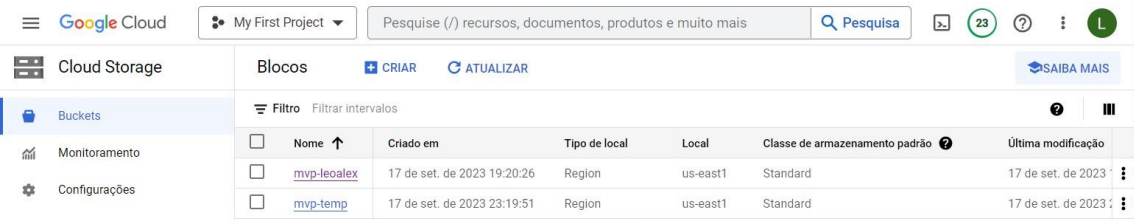
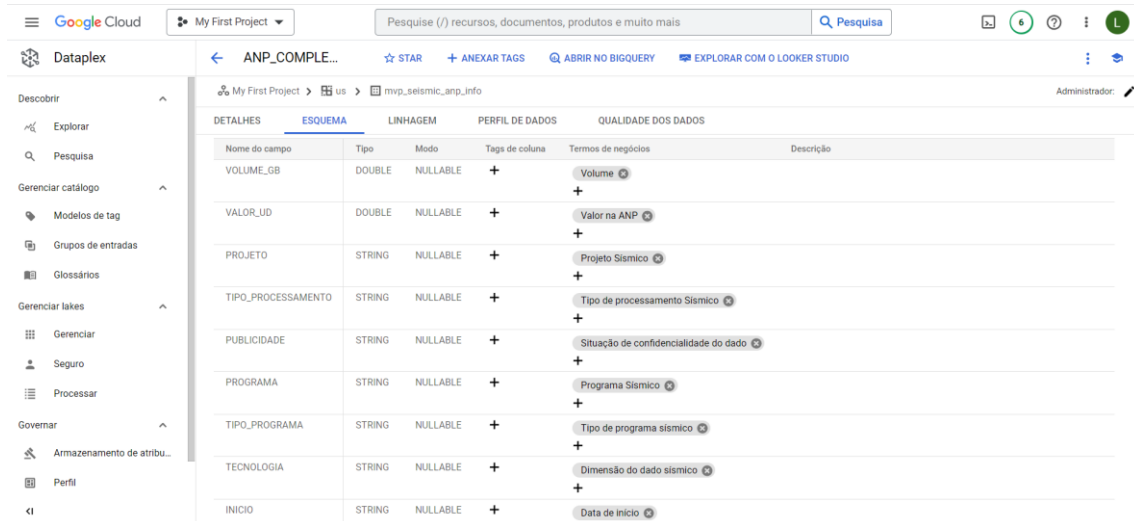


Figura 2 - Bucket do Cloud Storage da Google Cloud.

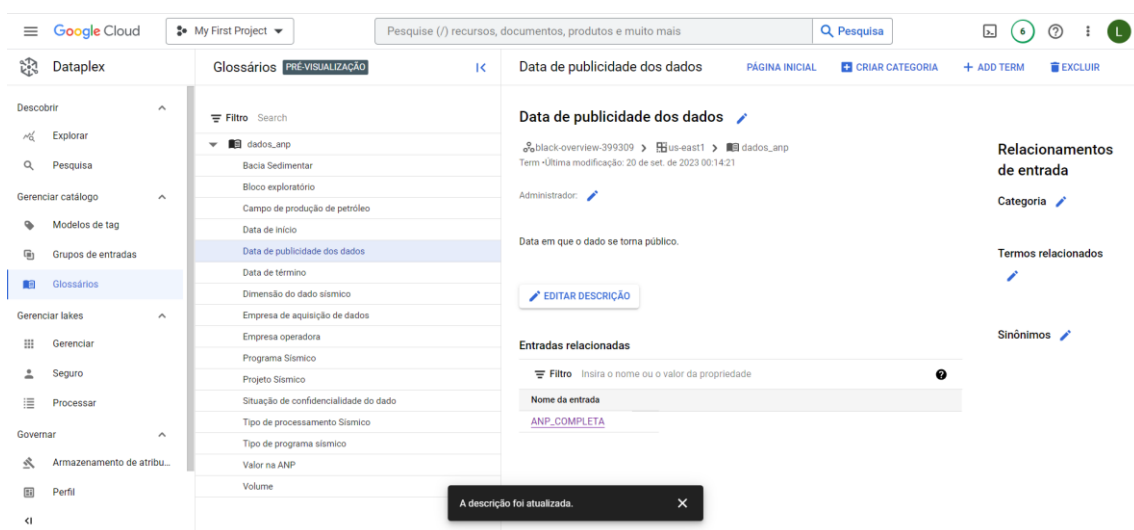
3. Modelagem

A modelagem do dado flat foi realizada na ferramenta Data Plex onde foram definidos cada termo da tabela e identificado o domínio dos dados.



Nome do campo	Tipo	Modo	Tags de coluna	Termos de negócios	Descrição
VOLUME_GB	DOUBLE	NULLABLE	+	Volume	
VALOR_UD	DOUBLE	NULLABLE	+	Valor na ANP	
PROJETO	STRING	NULLABLE	+	Projeto Sísmico	
TIPO_PROCESSAMENTO	STRING	NULLABLE	+	Tipo de processamento Sísmico	
PUBLICIDADE	STRING	NULLABLE	+	Situação de confidencialidade do dado	
PROGRAMA	STRING	NULLABLE	+	Programa Sísmico	
TIPO_PROGRAMA	STRING	NULLABLE	+	Tipo de programa sísmico	
TECNOLOGIA	STRING	NULLABLE	+	Dimensão do dado sísmico	
INICIO	STRING	NULLABLE	+	Data de início	

Figura 3 – Catálogo dos dados.



Nome da entrada
ANP_COMPLETA

Figura 4 - Realizando a descrição do atributo.

4. Carga

Foi criada a instância do Data Fusion para a carga dos dados.



Nome da instância	Ação	Edição	Região	Zona	Versão	Criptografia	Criada	Última atualização	Rótulo
mvp-la-anp	Ver instância	Basic	us-east1	-	6.9.2 (latest version)	Gerenciada pelo Google	17 de set, de 2023, 23:23:58	17 de set, de 2023, 23:34:02	-

Figura 5 - Instância criada.

Após diversas tentativas frustradas foi utilizado o DataPrep da Google para gerar um novo arquivo csv visando um melhor ajuste e facilitação da carga final no Bigquery.

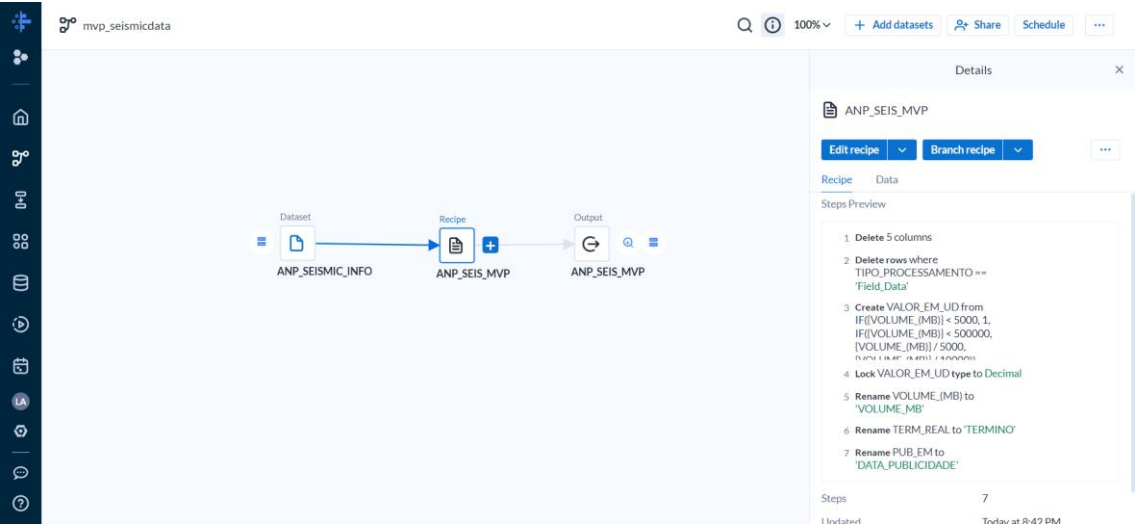


Figura 6 - Utilização do Dataprep para a melhor do csv e criação de uma coluna.

Após a criação do arquivo csv foi montado o fluxo de trabalho no Cloud Data Fusion. O dado era carregado através do GCS, posteriormente ajustado utilizando a ferramenta de transformação Wrangler e posteriormente ligado ao BigQuery.

The screenshot shows the Google Cloud Data Fusion Wrangler interface. The top bar includes 'Cloud Data Fusion | Wrangler' and navigation links for 'OPERATIONS', 'HUB', 'SYSTEM ADMIN', and 'Basic Edition'. Below the header, a tab for 'ANP_SEIS_MVP.csv' is active, showing a preview of the data with columns: VOLUME_MB, PROJETO, TIPO_PROCESSAMENTO, CONFIDENCIALIDADE, and PROGRAMA. The data is organized into a table with 10 rows. On the right, a 'Transformation steps (46)' panel is open, listing various transformations applied to the data, such as 'rename VOLUME_MB_VOLUME_MB', 'drop COD', 'find-and-replace CONFIDENCIALIDADE s/ig', 'drop CATEGORIA', 'find-and-replace TIPO_LEVANTAMENTO s/ig', 'find-and-replace TIPO_LEVANTAMENTO s/ig', 'find-and-replace TECNOLOGIA s/Sig', 'drop AUTORIZACAO', 'drop ATO_NORM', and 'find-and-replace TECNOLOGIA s/multiples/mul'.

	Double	String	String	String	String
	VOLUME_MB	PROJETO	TIPO_PROCESSAMENTO	CONFIDENCIALIDADE	PROGRAMA
1	43.0576	0225_PARANA_57	MIG_FIN	Publico	0225_PARANA_57
2		0275_2D_SPEC_BFZ_PH1_RM13	Field_Data	Publico	0275_2D_SPEC_BFZ...
3	65.253	0026_POTIGUAR_39	MIG_FIN	Publico	0026_POTIGUAR_39
4	1.2129	R0011_E_SANTO	MIG_FIN	Publico	R0011_E_SANTO
5		0268_4C-VIOLAO	Field_Data	Publico	0268_4C-VIOLAO
6		0268_4C-CACHALOTEJUBARTE	Field_Data	Publico	0268_4C-CACHALOTE
7	82.5664	0284_2DOBC1_CES134BCE5	MIG_FIN	Publico	0284_2DOBC1_CES13
8	301.3536	0026_2D_ITANHAIUA_JUMA	MIG_FIN	Publico	0026_2D_ITANHAIUA...
9	2148.2499	R0015_FOZ_DO_AMAZONAS_PHASE2	MIG_PSTM	Publico	R0015_FOZ_DO_AMA...
10	33824.2549	R0003 GRAND FOZ DO AMAZONAS	MIG_FIN	Publico	R0003 GRAND FOZ...

Figura 7 - Transformação no Wrangler

No Wrangler foram excluídos alguns atributos considerados desnecessários e ajustados outros itens com problemas de visualização.

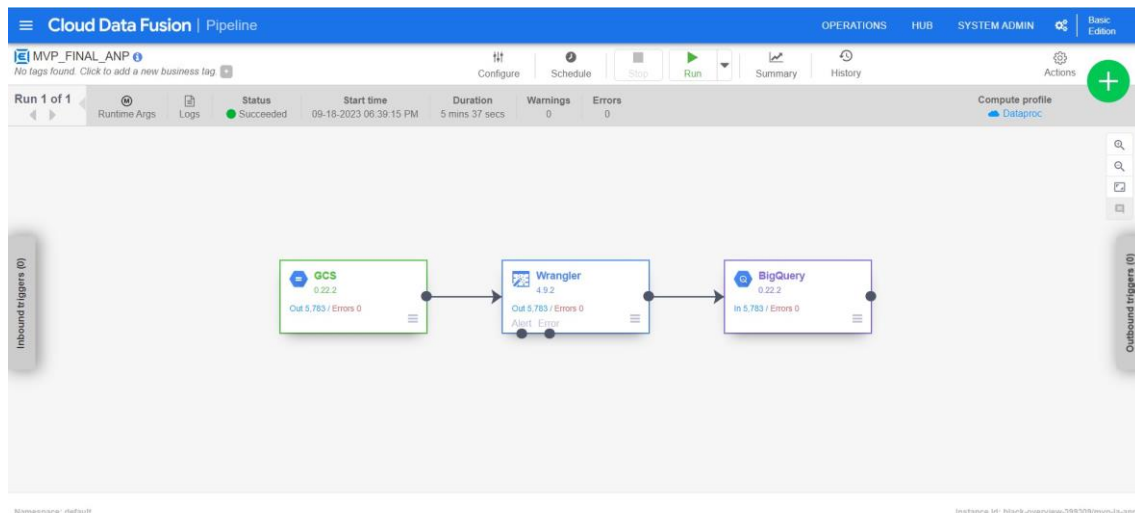


Figura 8 - Fluxo criado no Cloud Data Fusion e bem-sucedido no BigQuery.

5. Análise

a. Qualidade dos Dados

A base de dados da ANP apresenta ausência de informações em todos os atributos com exceção do atributo “PROGRAMA” e da “PUBLICIDADE”. Esses problemas não afetam todos os objetivos do MVP, porém deixam algumas respostas incompletas. O atributo “VOLUME_GB”, que trata do volume dos dados sísmicos, apresenta grande quantidade de dados sem informações (NULL), logo o cálculo de volume é afetado e consequentemente o cálculo total de valor. As informações do atributo “BACIA” também apresentam ausência de dados levando a não ser possível calcular o valor de todos os dados das bacias. A “PUBLICIDADE” é um dos poucos atributos totalmente preenchidos, sendo este de fato um atributo muito importante por tratar da confidencialidade do dado.

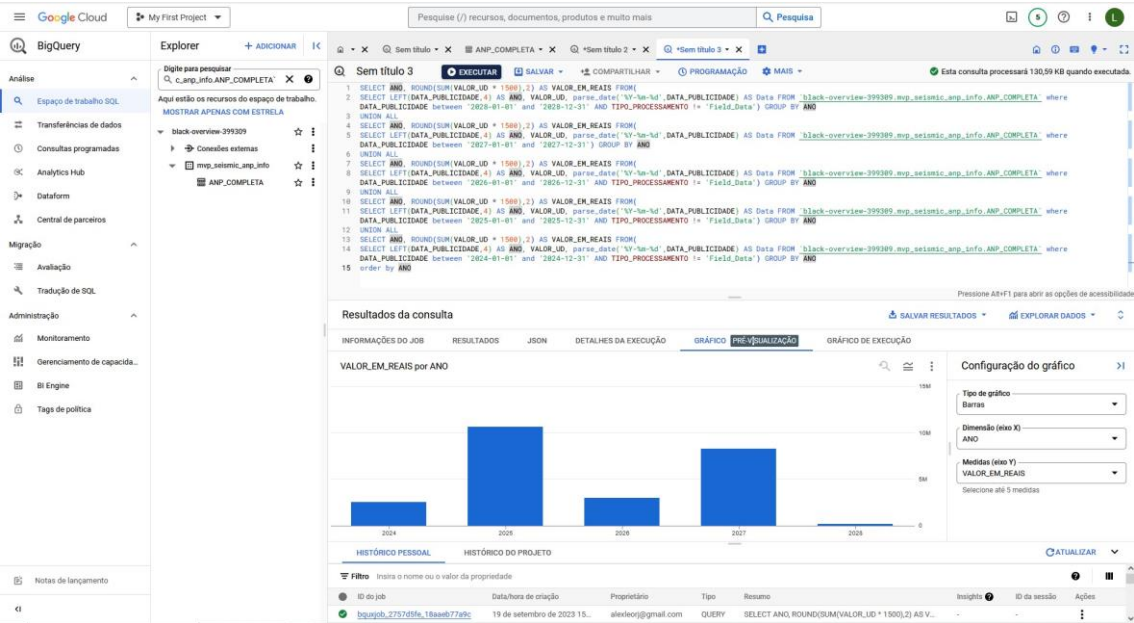
Por se tratar de um problema interno da Agência a única maneira de resolver certas informações seria pesquisar afundo o histórico da exploração de petróleo no Brasil, mesmo assim, acredito ser um problema de difícil resolução.

b. Solução do Problema

Quanto uma empresa gastará ao longo dos próximos anos comprando os dados que a ANP para manter seu banco atualizado?

É possível prever através dos dados da planilha os gastos anuais que uma empresa deve ter para manter o banco similar ao da ANP. Com isso o planejamento da reserva financeira para a compra de dados fica facilitado e pode ser planejado de maneira mais concreta pois dos anos mais recentes as informações dos dados são mais confiáveis e completas.

A consulta sql foi utilizada para definir os programas que ficarão públicos durante o ano desejado e o valor em reais da compra de todos os dados disponíveis naquele ano.



ANO	VALOR_EM_REAIS
2024	253998.02
2025	10677097.35
2026	299794.76
2027	8288006.18
2028	185603.16

Figura 10 - Tabela com valores anuais no BigQuery.

É possível para uma empresa igualar seu banco de dados com o da ANP?

De acordo com a informação de valores referentes a todos os dados da ANP não ser muito precisa, devido a falta de informação em muitos projetos sísmicos a pergunta fica sem resposta.

Uma empresa chegando ao Brasil hoje, é mais econômico comprar os dados da ANP ou fazer seu próprio levantamento geofísico?

Tendo o conhecimento do valor de uma aquisição sísmica por ter trabalhado em 3 projetos terrestres e analisando o valor encontrado na pesquisa sql no BigQuery, é possível afirmar que é mais econômico comprar dados com a ANP do que os adquirir.

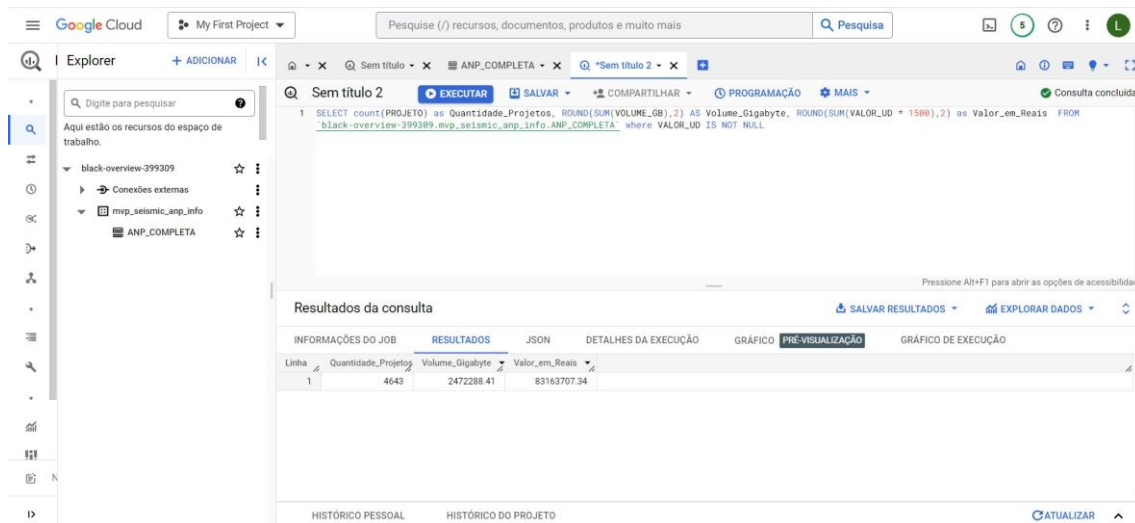


Figura 11 - Valor total dos dados disponíveis na ANP e com informação de volume divulgada.

De acordo com a figura 11, o valor de 4643 projetos ultrapassa 83 milhões de reais na ANP, uma aquisição de um projeto terrestre em 2017 custou em torno de 50 milhões de reais. A figura 12 mostra que se tratando apenas de dados públicos, que são os dados que podem realmente ser comprados no dia de hoje o valor diminui bastante, mesmo se tratando de processamentos de maior custo como dados com tipo de processamento PSDM.

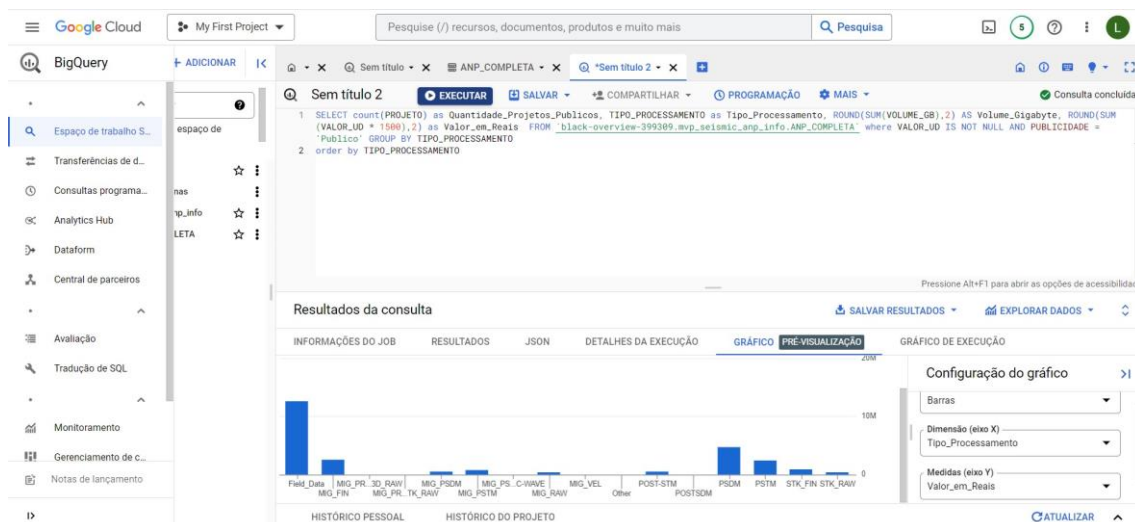


Figura 12 - Valor dos dados por processamento.

A ANP negocia esses dados a um valor coerente? Dentro do valor gasto para adquirir dados sísmico, o valor cobrado pela agência é de fato compatível com o que foi gasto para criar esses dados?

A ANP foi criada próximo ao ano 2000, naquela época os dados não eram valorizados como atualmente, logo muitos dados antigos da ANP não possuem uma qualidade interessante para quem compra esses dados. Se tratando de dados mais recentes a história muda bastante com toda a valorização e investimento em tecnologia e armazenamento feito na Agência. O valor dos dados públicos na ANP chega perto da casa do 30 milhões, como demonstrado na figura 13. A ANP não leva em conta a idade dos dados, nem a tecnologia aplicada no processamento. O preço cobrado pelos dados é igual sendo eles recentes ou não. Como um dado leva em média 10 anos para se tornar público na ANP, acredito que o preço desses 10 anos de teórico atraso faz com que os dados do governo fiquem mais barato, porém acredito que deveria ser levando em conta a idade do dado para balancear o valor aplicado.