**Pontifícia Universidade Católica do Rio de Janeiro - PUC - RJ**

**Ano: 2023**

**Sprint: Machine Learning & Analytics (40530010056\_20230\_01)**

**MVP - Machine Learning**

**Aluno:** ALAN ROLEMBERG ELIODORIO

**0. Objetivo do trabalho**

O comércio eletrônico vem crescendo a cada ano no mundo, e é um dos principais casos de uso para análise de dados, tanto no quesito de análise histórica em DWs como no quesito de análise preditiva, através de algoritmos de machine Learning.

Portanto, neste trabalho iremos realizar uma busca de uma base de dados de vendas via comércio eletrônico, buscando efetuar análises e responder as perguntas abaixo:

- Qual o incremento de vendas de produtos de um ano para o outro?

- Quais os produtos que mais venderam nos últimos anos?

- Quais os produtos que caíram em número de vendas nos últimos anos?

- Quais os métodos de pagamentos mais usados por consumidores jovens(<30)? E quais os métodos de pagamentos mais usados por consumidores idosos (>65)? E quais os métodos de pagamentos mais usados pelos consumidores em geral?

- Qual o índice de satisfação médio do consumidor do Nordeste do Brasil? E qual o índice de satisfação médio do consumidor do Sul do Brasil?

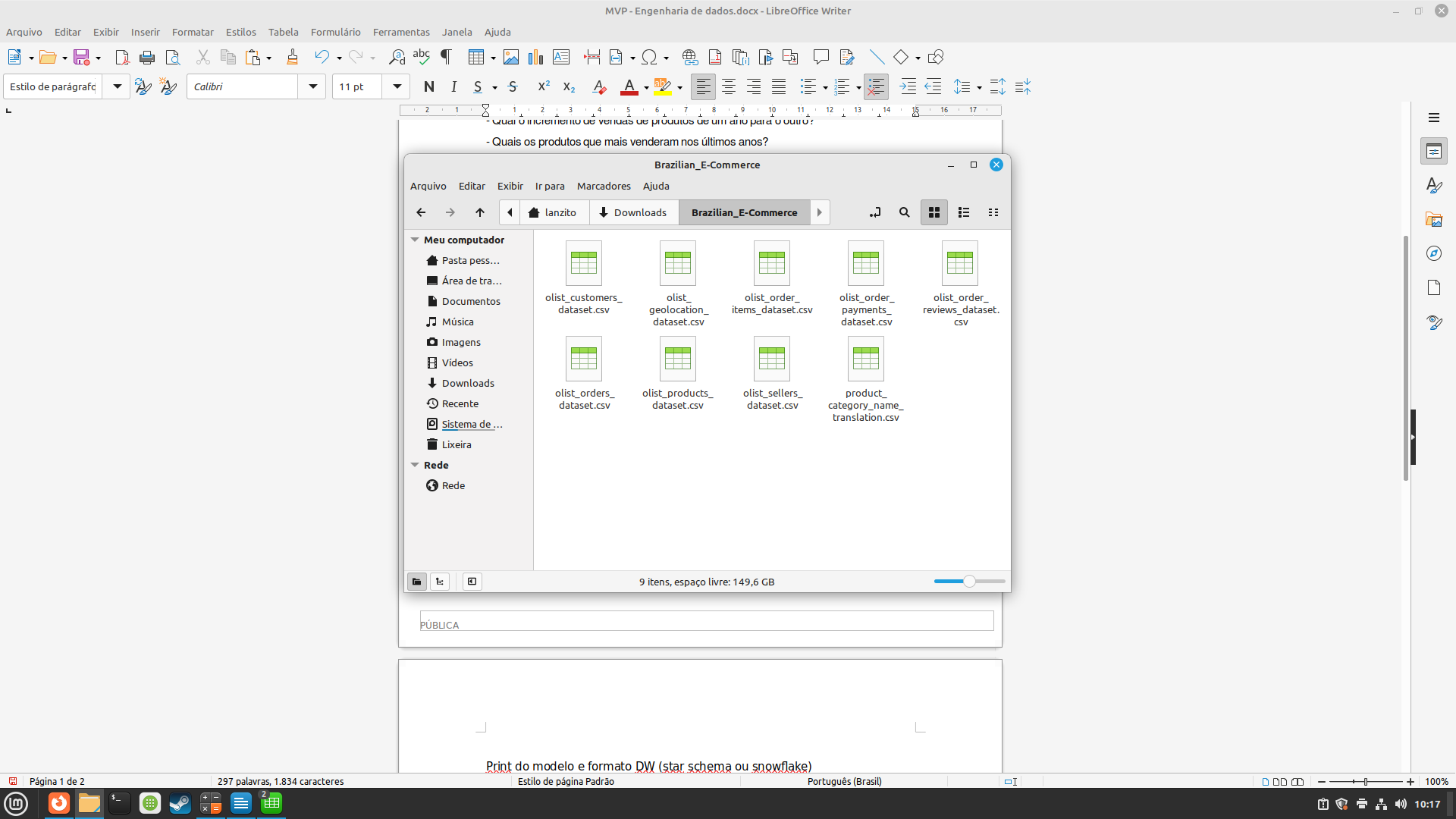
**1. Busca pelos dados**

A base de dados escolhida foi a Brazilian E-Commerce Public Dataset extraída do Kaggle. Este banco de dados foi disponibilizado pela olist.com, que é uma empresa que comercializa “sites para comércio eletrônico”.

URL: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?resource=download

**2 . Coleta**

Conseguimos resgatar um conjunto de 9 arquivos conforme imagem abaixo:

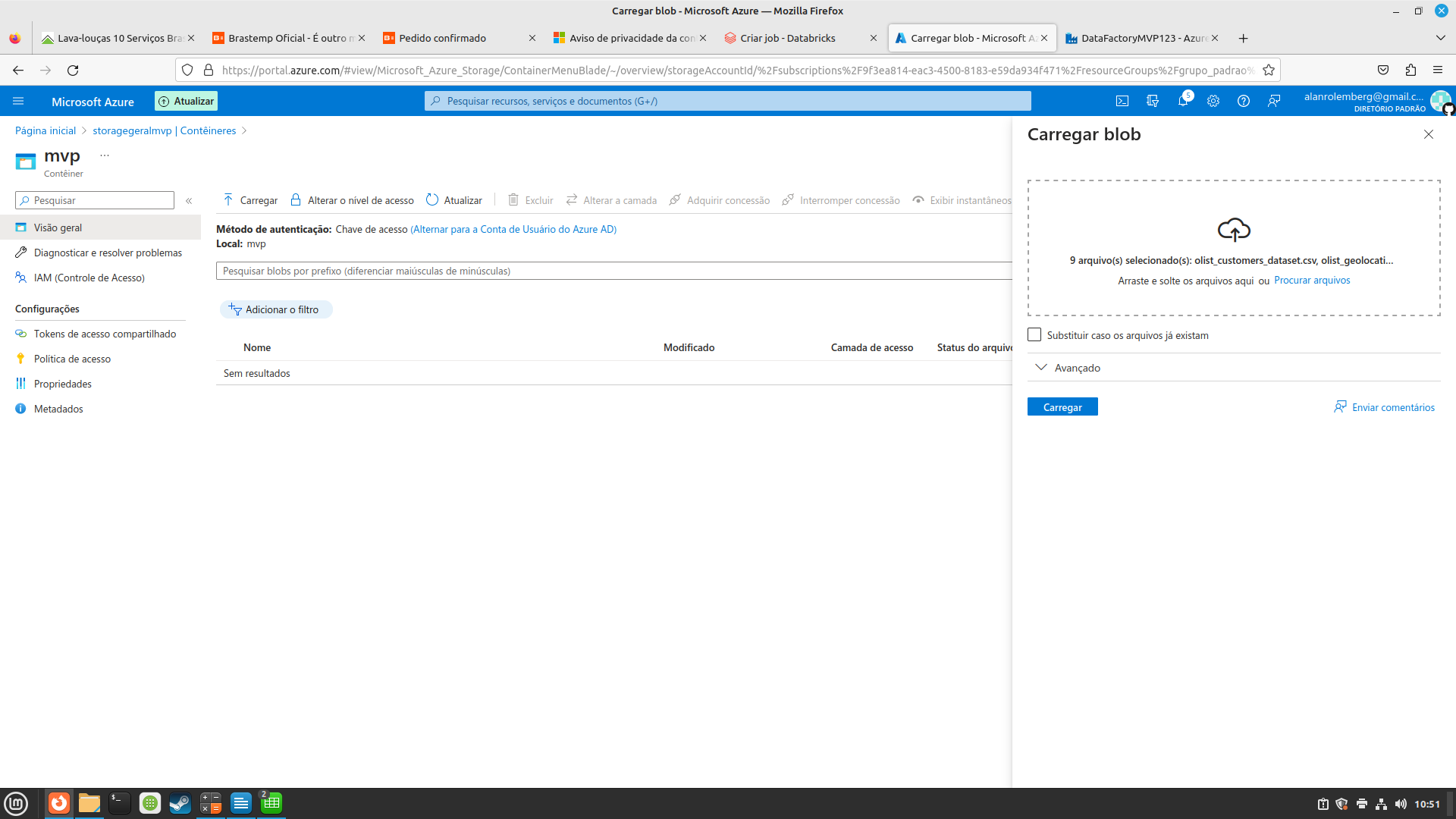


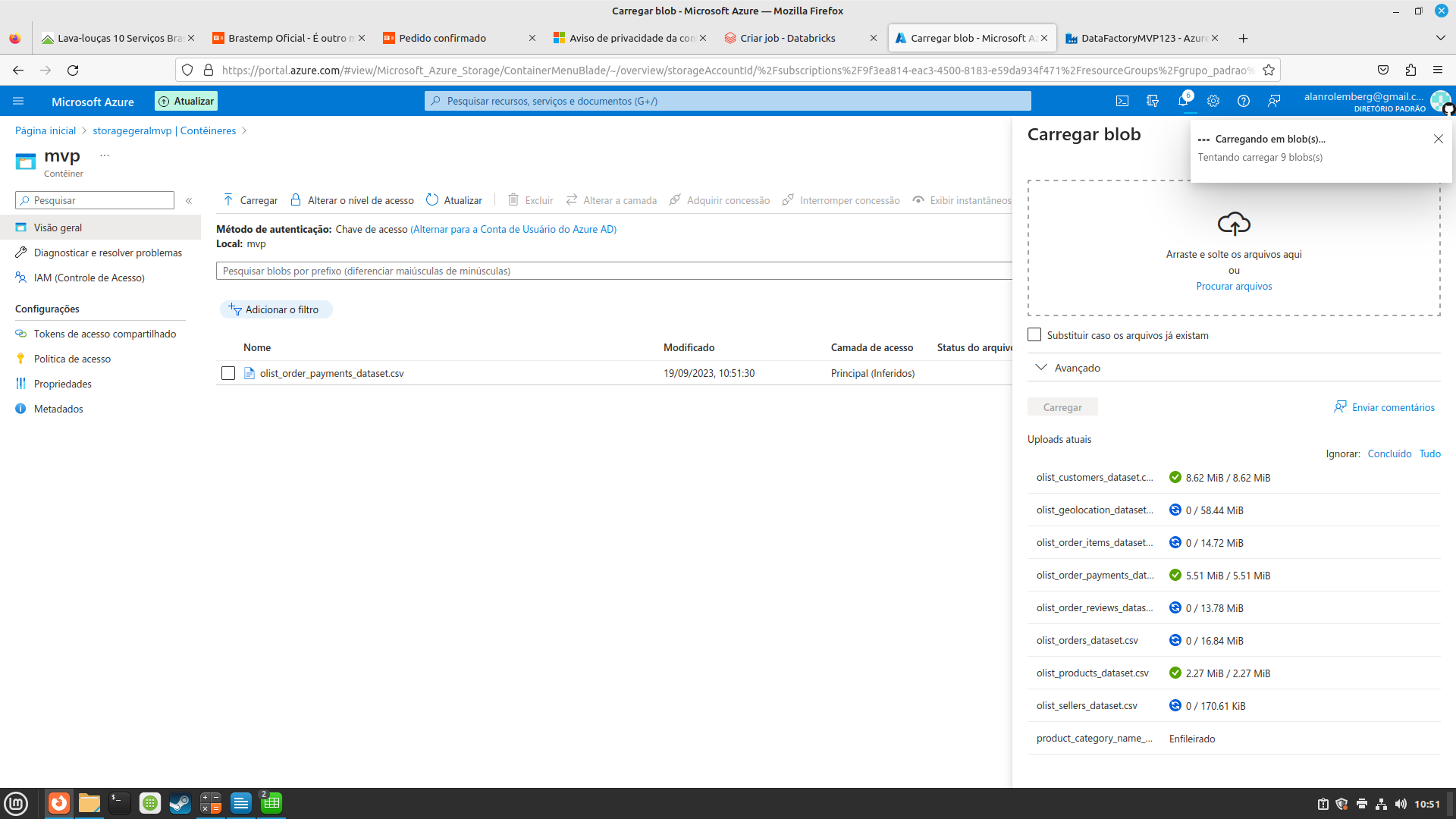
Esses arquivos representam tabelas normalizadas que contêm as entidades abaixo:

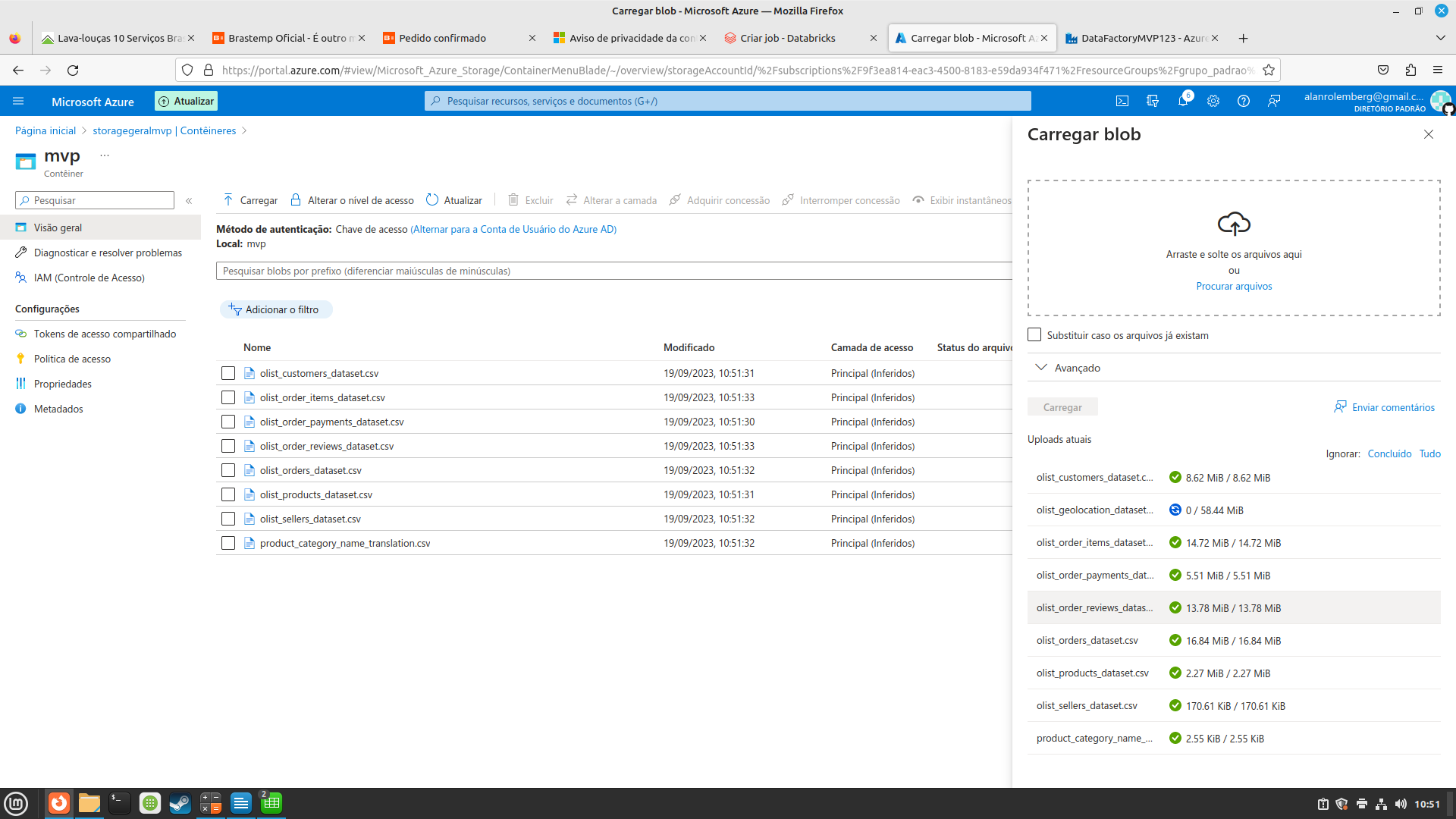
* Vendas
* Itens vendidos
* Produtos
* Clientes
* Vendedores
* Pagamento de pedidos
* Avaliação de pedidos
* Tradução do nome da categoria
* Geolocalização

Esses arquivos serão carregados no Azure Databricks, e armazenados numa base de dados normalizada SQL Server.

O primeiro passo que executamos foi o carregamento dos arquivos no Azure Blob Storage, para fins de uso dentro dos recursos da nuvem Azure.

A tela a seguir mostra o carregamento dos arquivos:



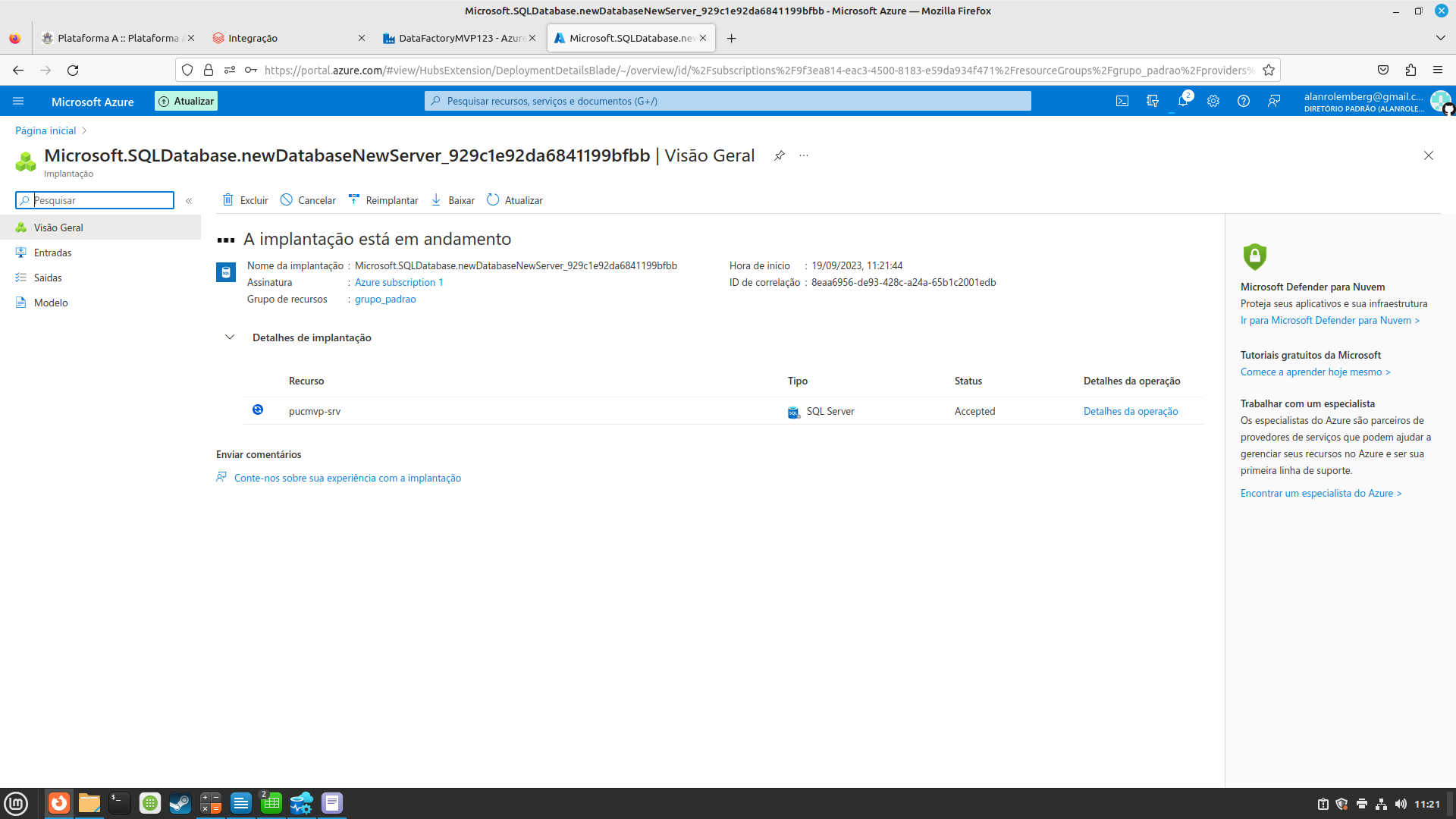


O segundo passo que executamos foi a criação da base de dados SQL Server que será usada como repositório inicial dos dados:

**Nome do recurso de banco de dados:** pucmvp-srv/pucmvp

**Nome do recurso de servidor:** pucmvp-srv

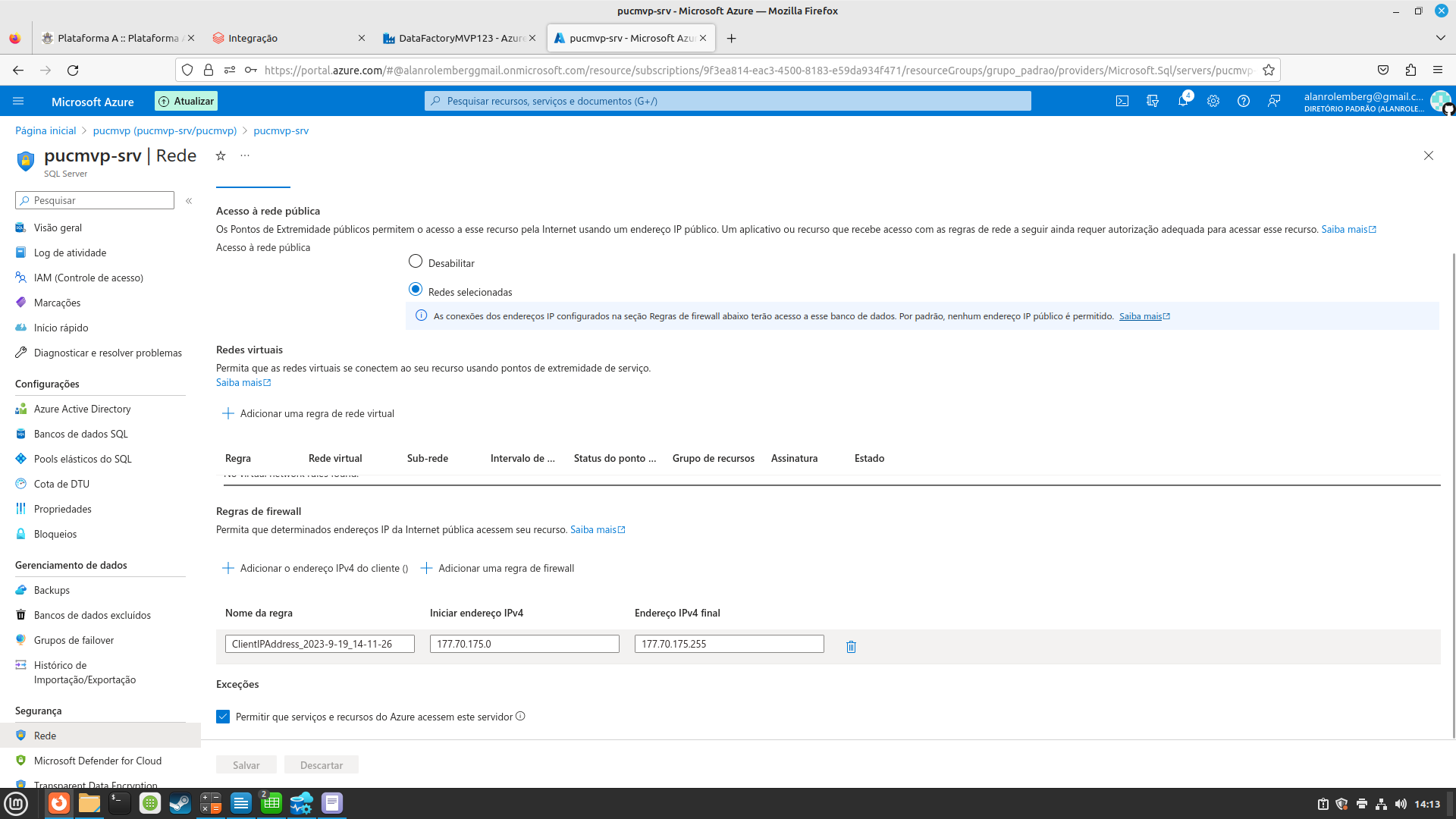
**Hostname para acesso público/privado:** pucmvp-srv.database.windows.net

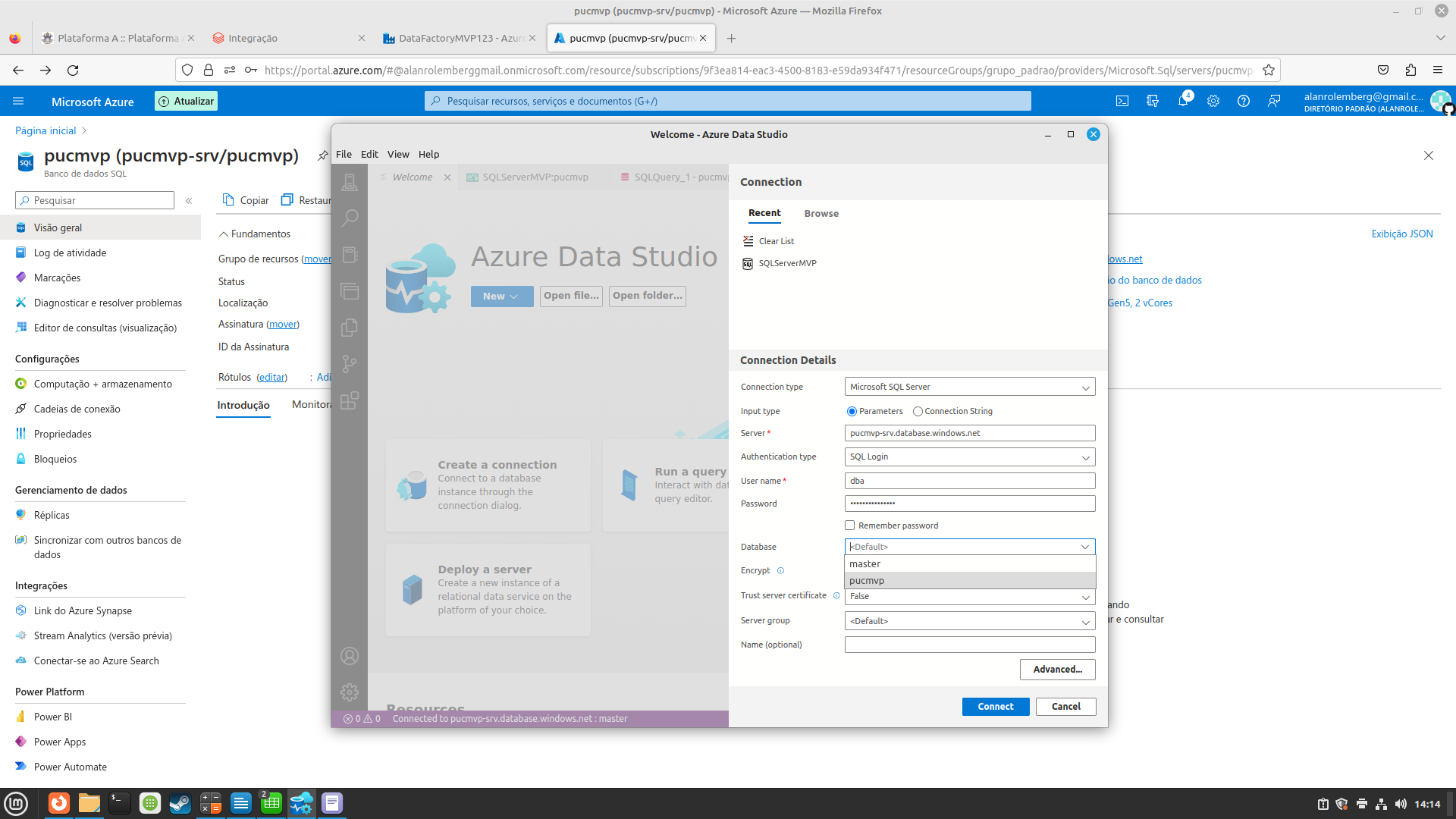


O passo seguinte foi a configuração do controle de acesso.

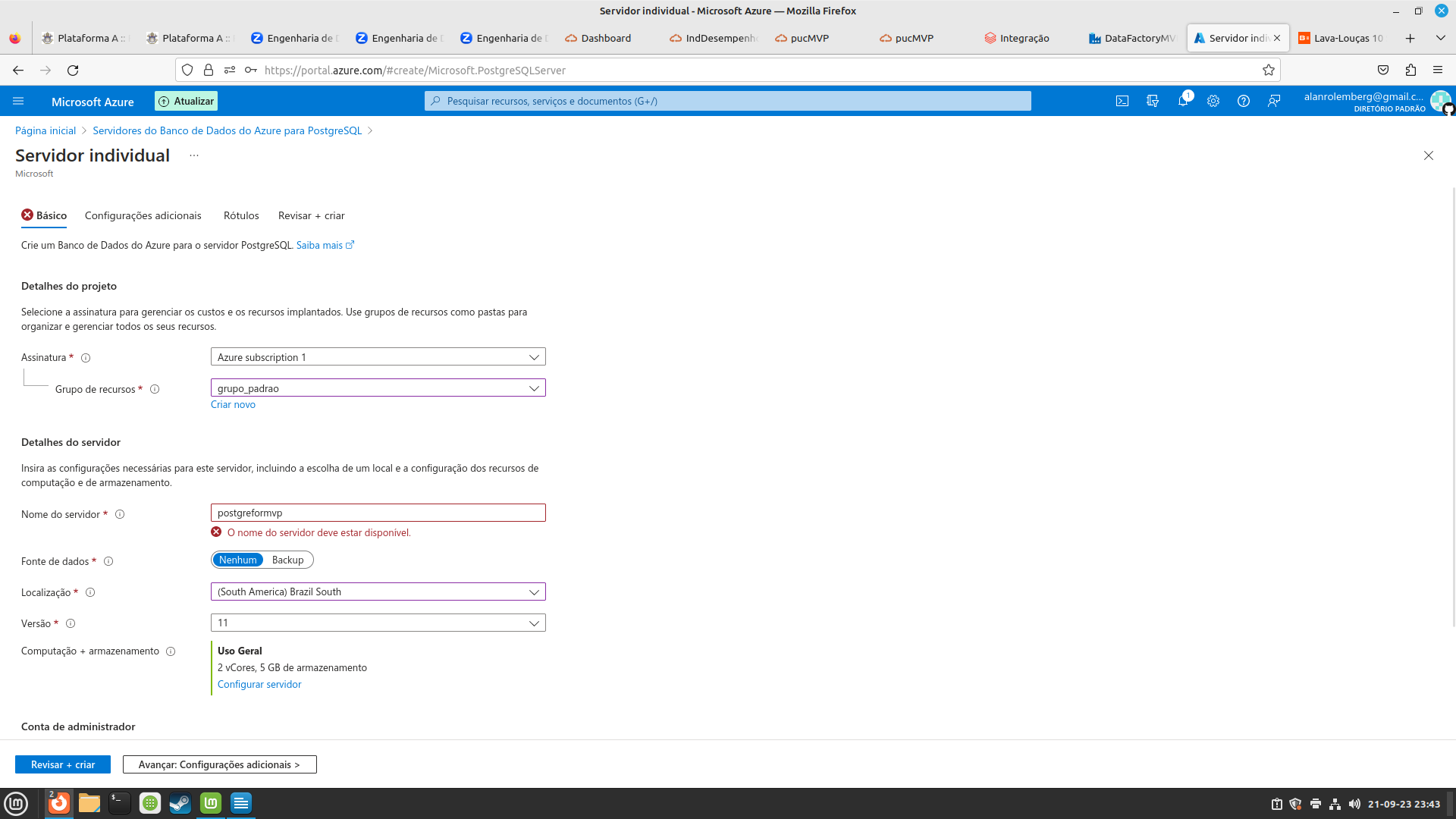
As regras de firewall foram abertas para o range de Ips 177.70.175.\*, pertencentes ao meu provedor de internet.

Foi marcada também a opção **Permitir que serviços e recursos do Azure acessem este servidor** para garantir que a ingestão de dados ou processo ETL consigam ler os dados através das ferramentas da nuvem Azure.

O passo seguinte foi o teste de acesso ao banco de dados através do Azure Data Studio:



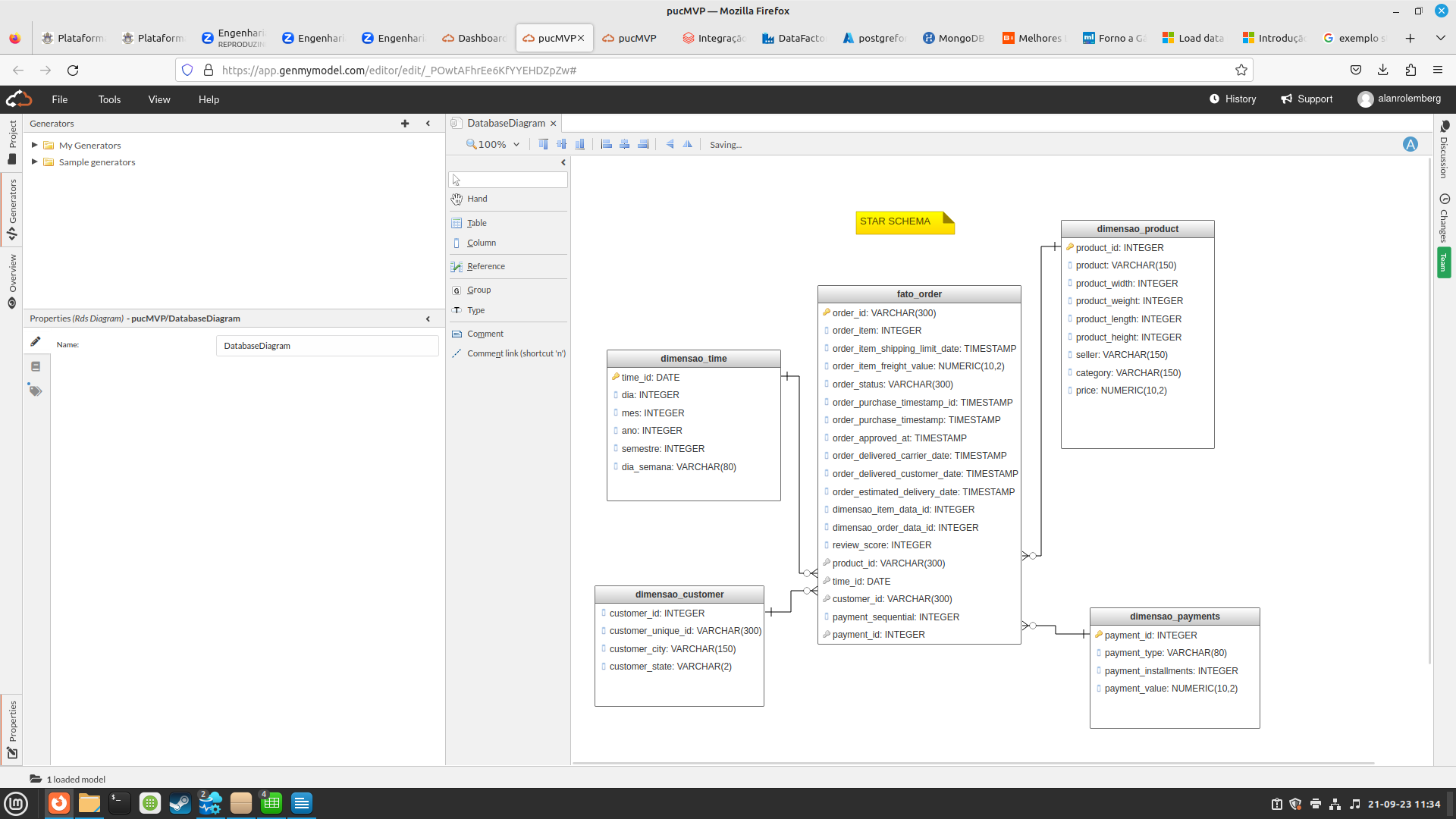
Após este passo, notamos que a cobrança por uma base de dados SQL Server estava alta. Então eliminamos o recurso e optamos por criar uma nova base de dados PostgreSQL:



**3 . Modelagem**

Idealizamos uma modelagem alvo para representar o datawarehouse que precisamos para responder nossas perguntas. O formato escolhido foi star schema, no qual teremos apenas uma tabela fato, que terá os dados das vendas (orders) que iremos analisar. Esse modelo foi elaborado através da ferramenta genmymodel.com .

Abaixo está a imagem do modelo que desenhamos:



Precisamos agora materializar este modelo em um ambiente de datawarehouse numa ferramenta de nuvem.

Para isto, utilizaremos o Azure Data Factory

Descritivo com o catálogo de dados

Print do catálogo de dados carregado na ferramenta de nuvem

Descritivo com os domínios dos campos: valor max e min e tabela de possibilidades para dados categóricos

4. Carga

Descrever a carga feita através do Azure Data Factory

5. Análise

        a. Qualidade de dados

          b. Solução do problema