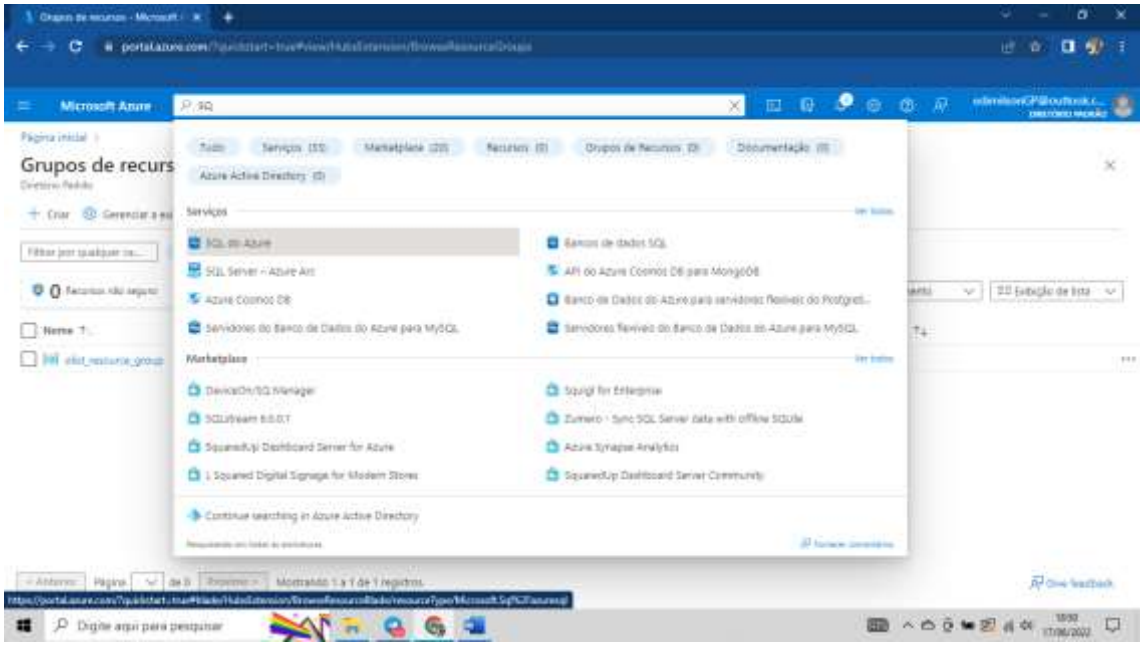
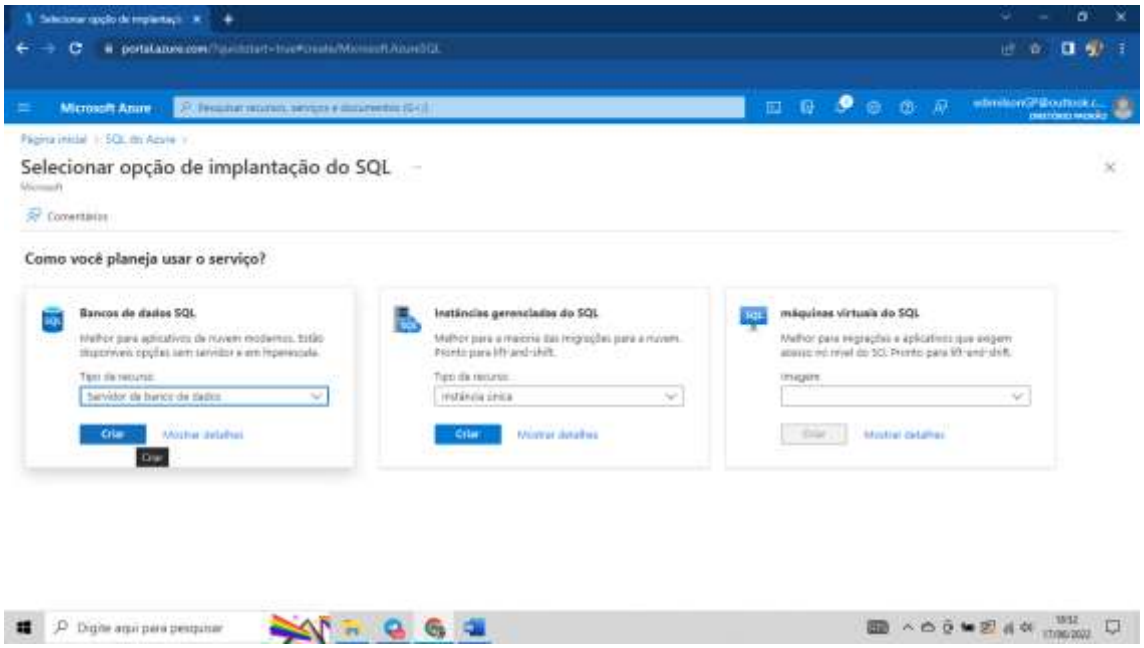
CRIAR BANCO DE DADOS SQL DO AZURE



SELECIONAR EM BANCO DE DADOS SQL

Servidor de banco de dados

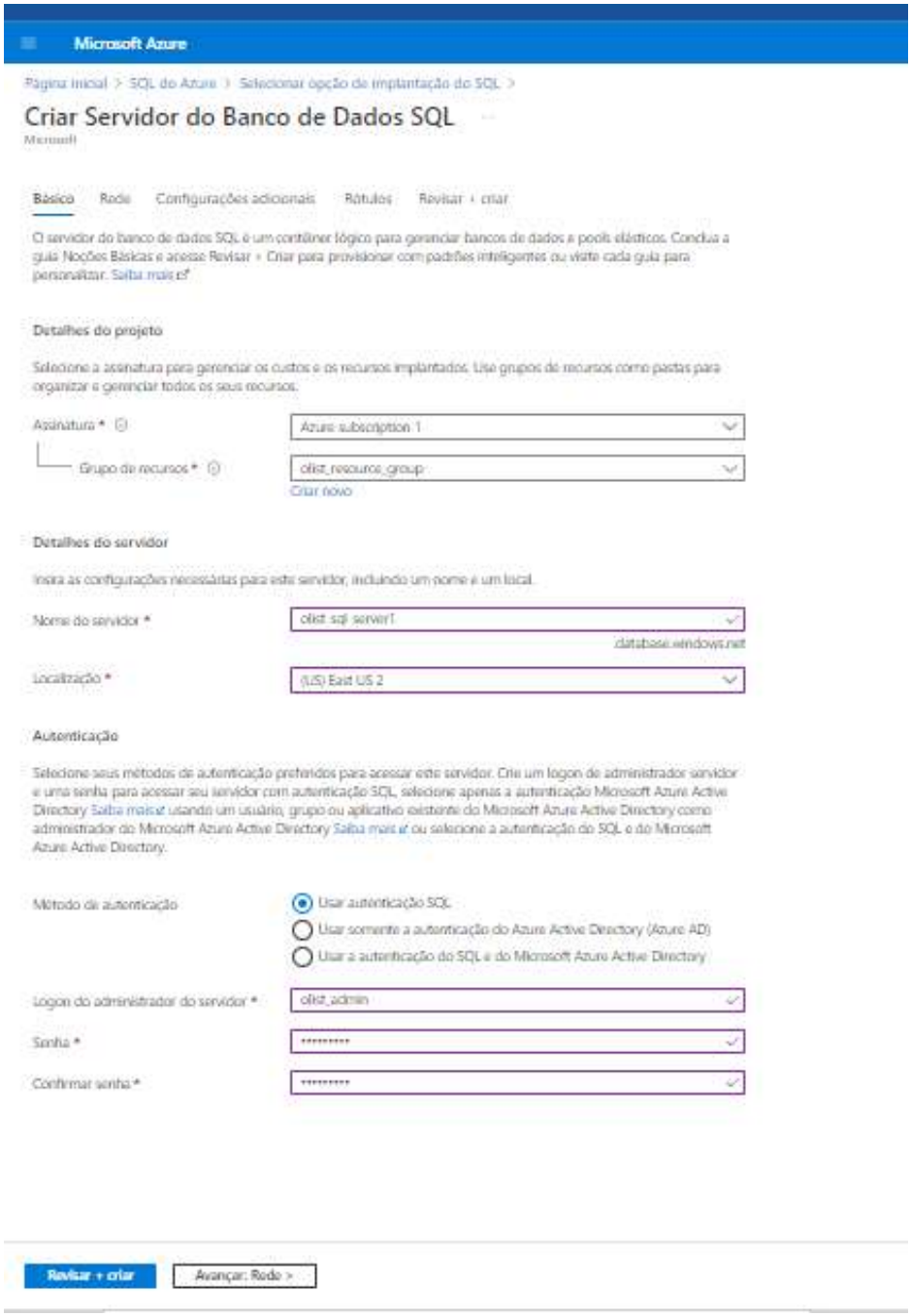


Nome do servidor: olist-sql-server1 .database.windows.net

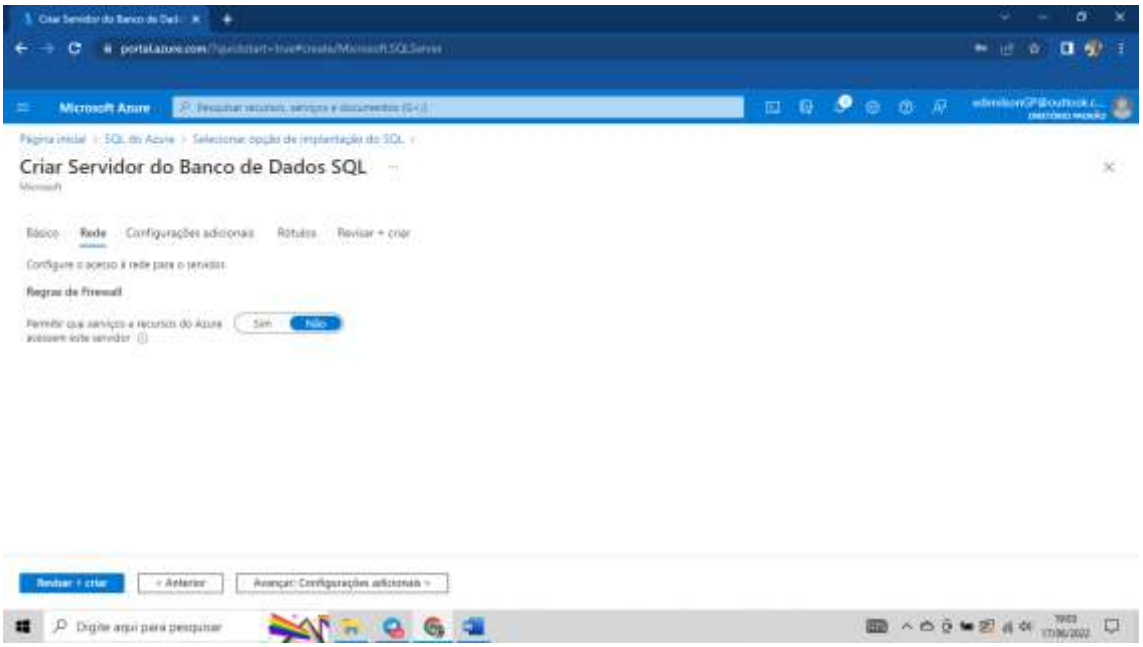
Logon do administrador do servidor: olist_admin

Senha:******

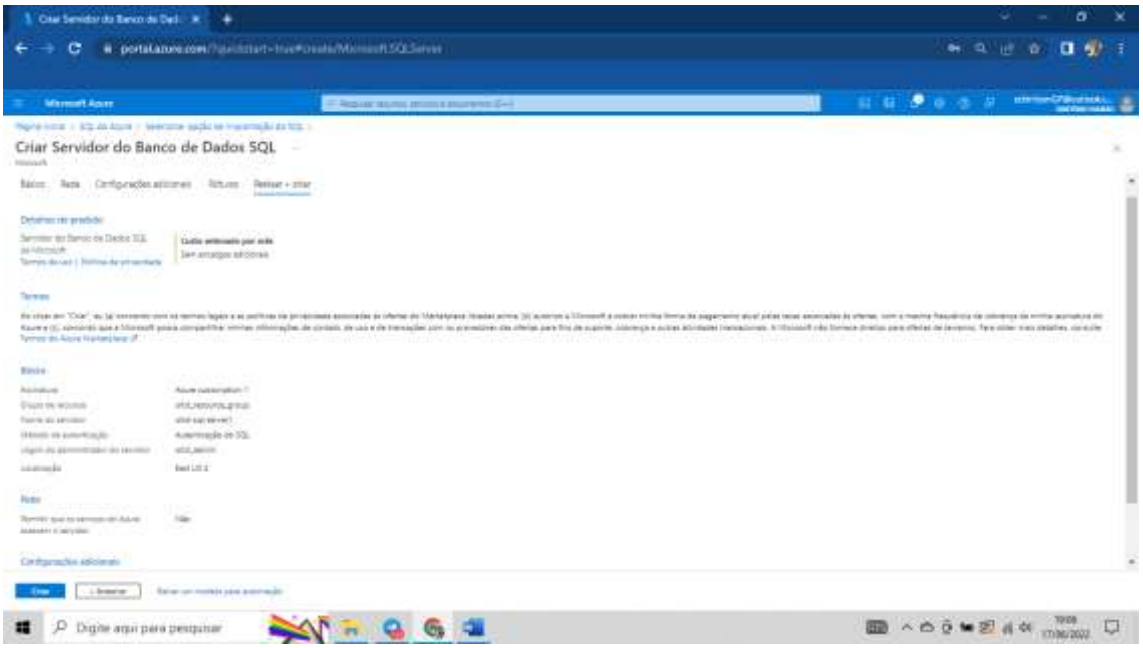
Clicar em AVANÇAR



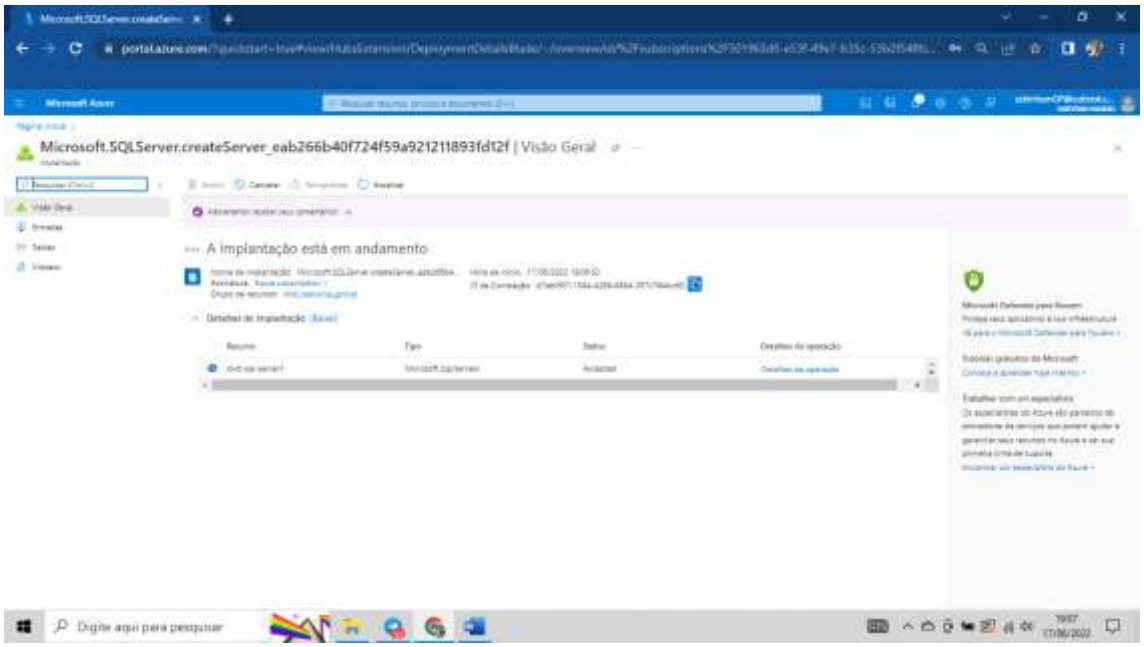
AQUI SÓ MANTER E CLICAR EM AVANÇAR



CLICAR EM CRIAR

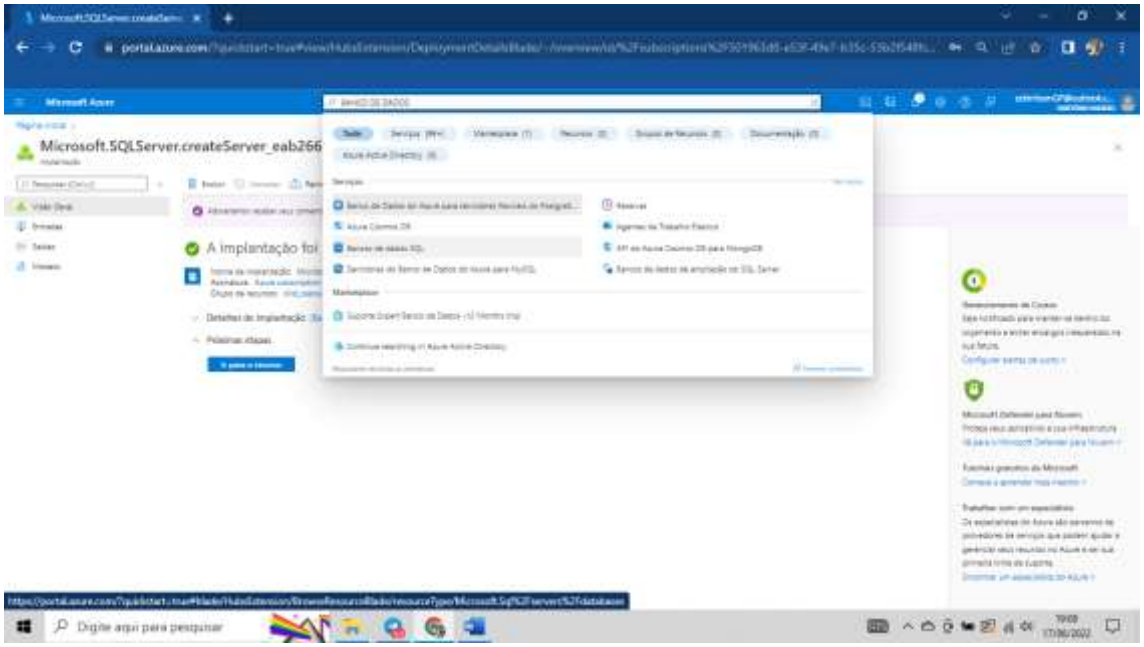


EM ANDAMENTO

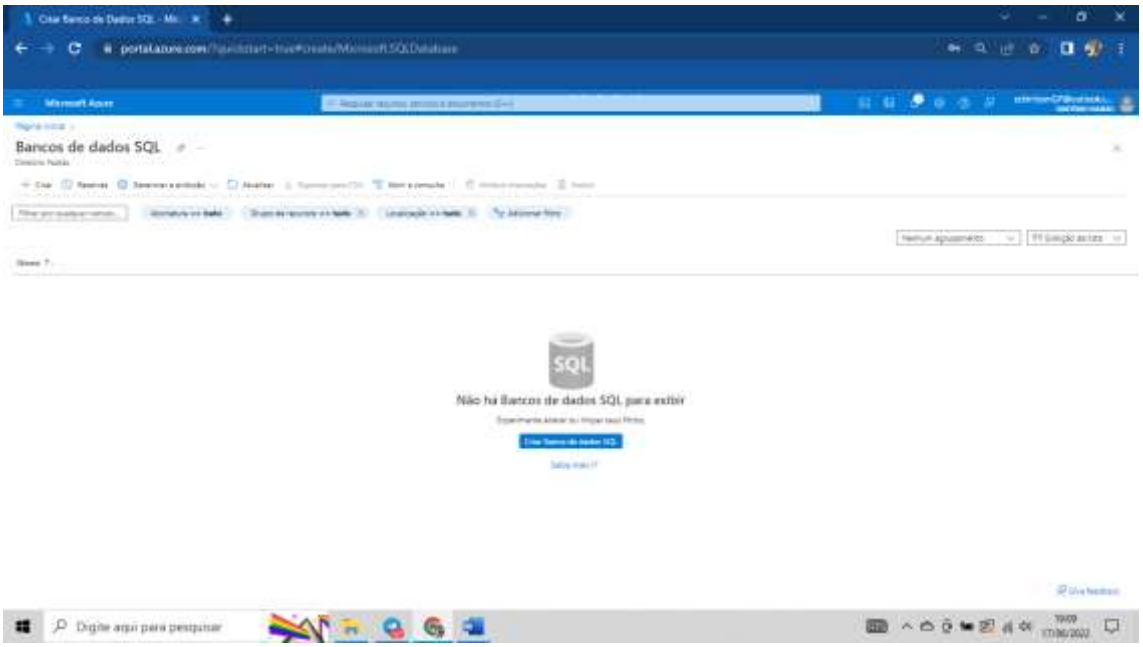


CONCLUIDA

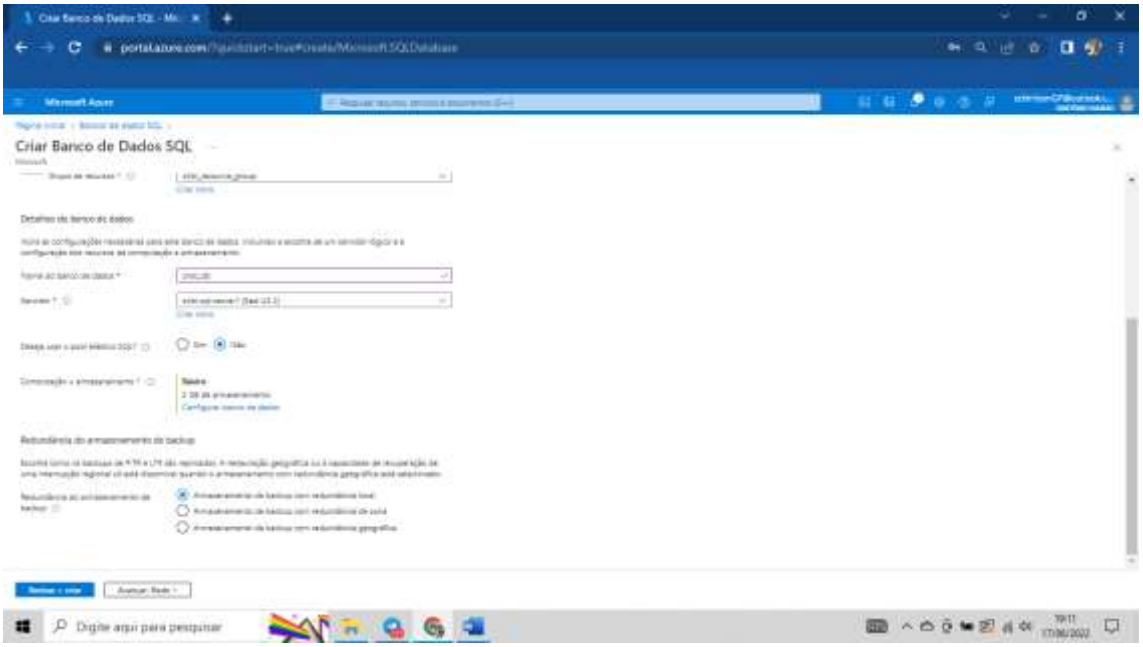
BANCO DE DADOS SQL



CRIAR

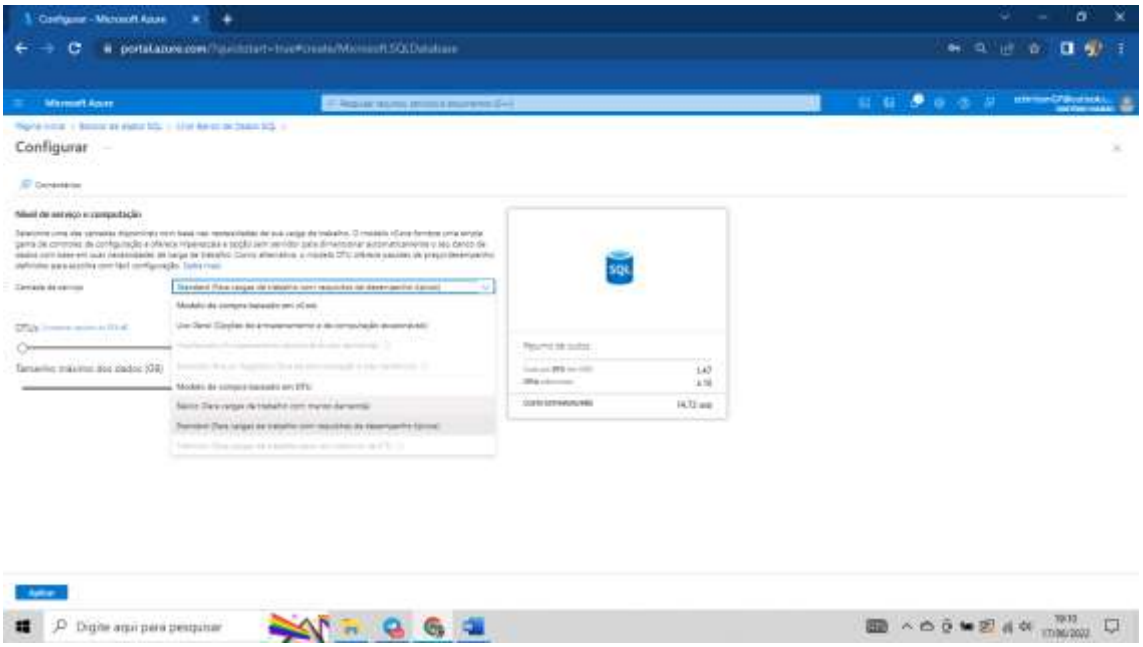


Nome do banco de dados: olist_db

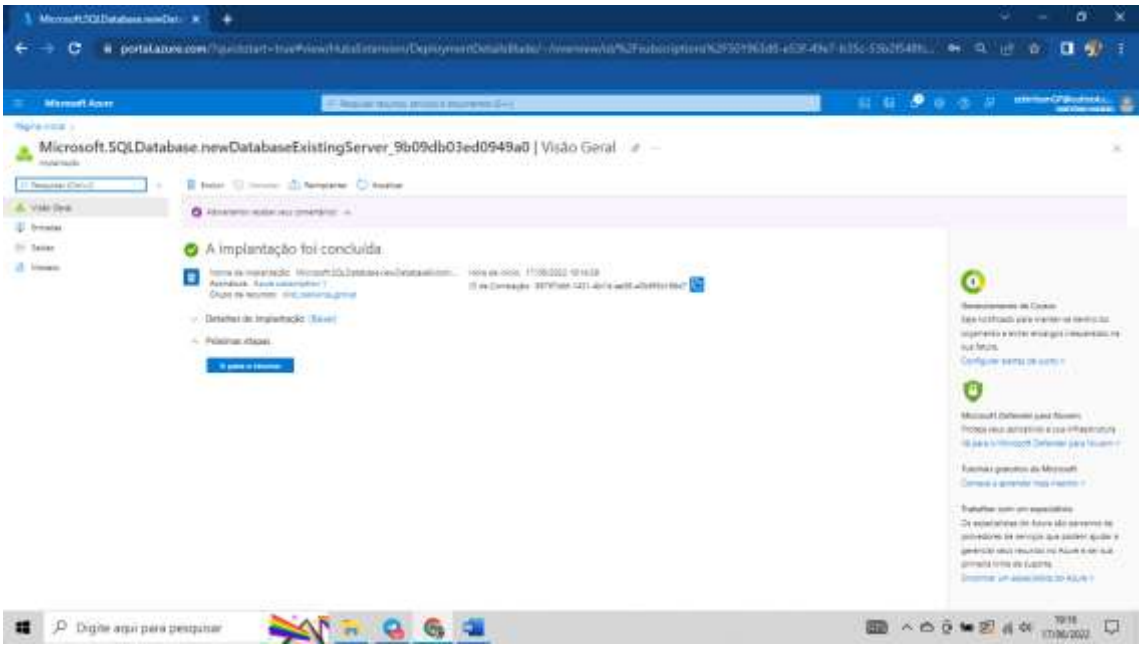


CONFIGURAÇÃO

AVANÇAR ATÉ O FINAL E CLICAR EM CRIAR



CRIADO O DATA BASE

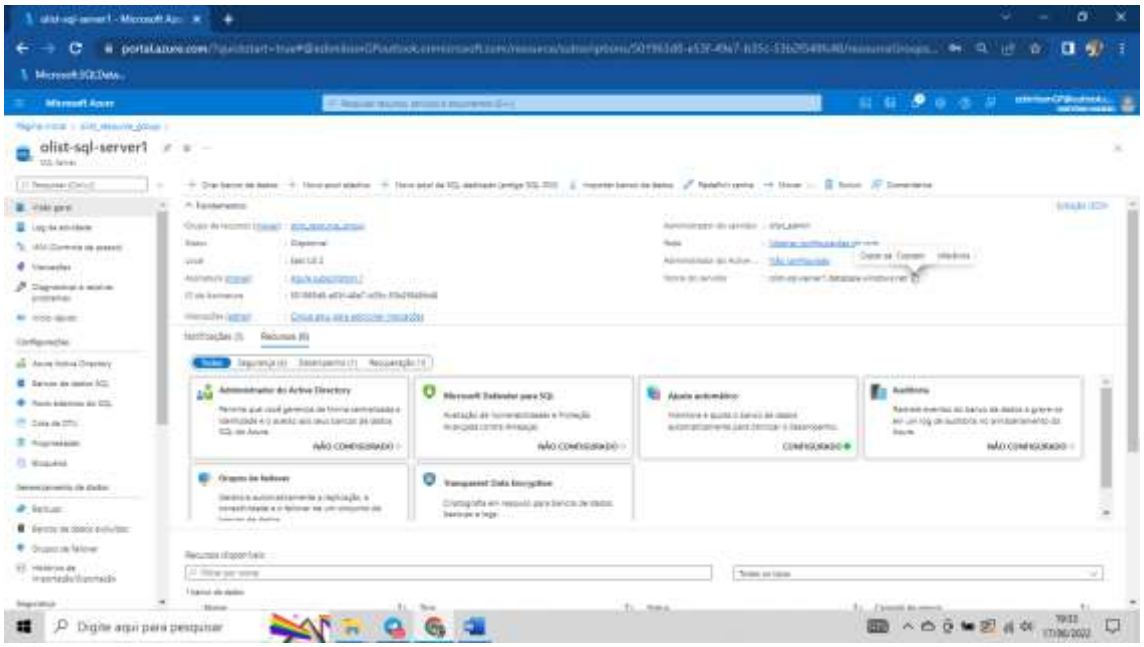


AGORA CARREGAR OS DADOS DENTRO DO DATA BASE

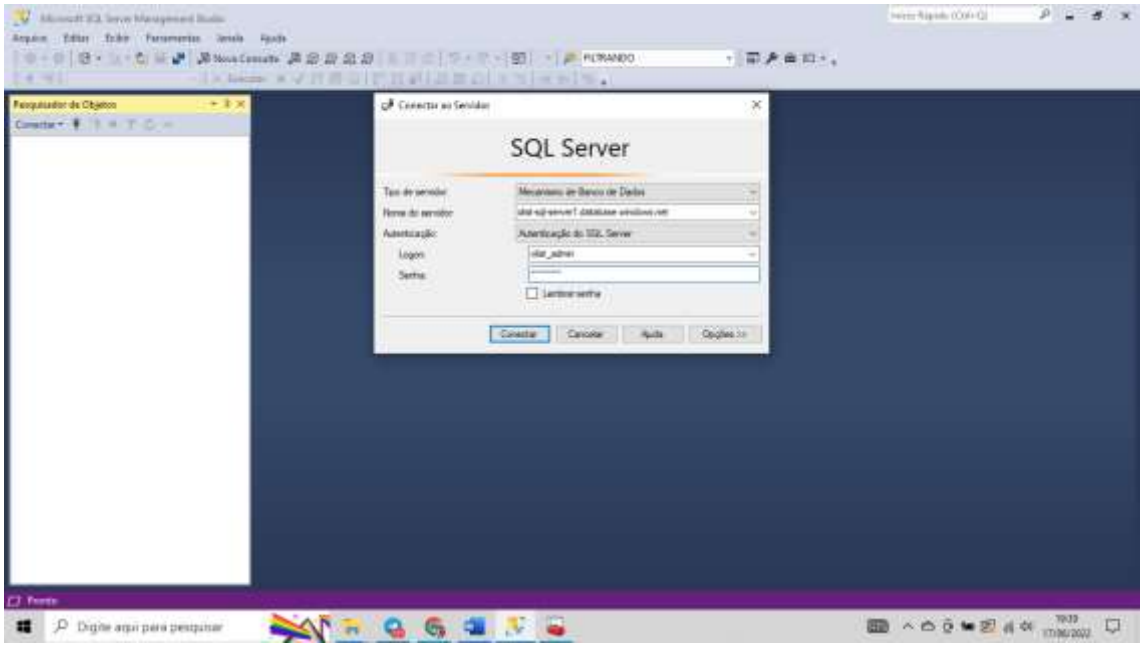
Link do dataset no kaggle: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Copiar Nome do servidor: olist-sql-server1.database.windows.net

Para fazer a conexão

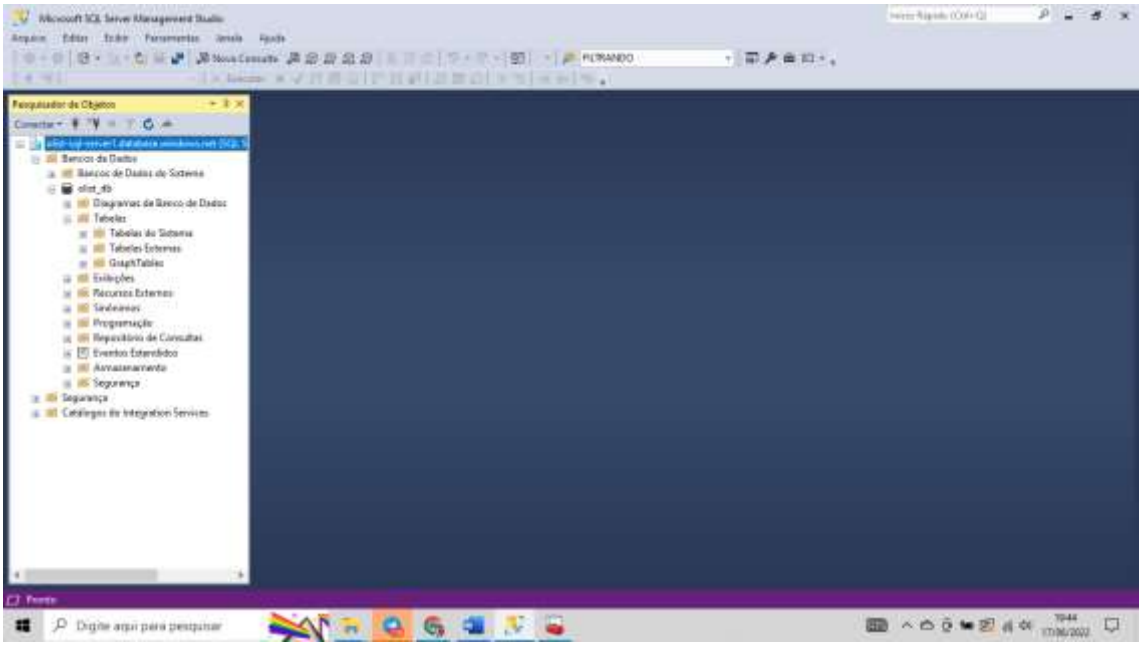


Inserindo os dados de conexão no SQL SERVER

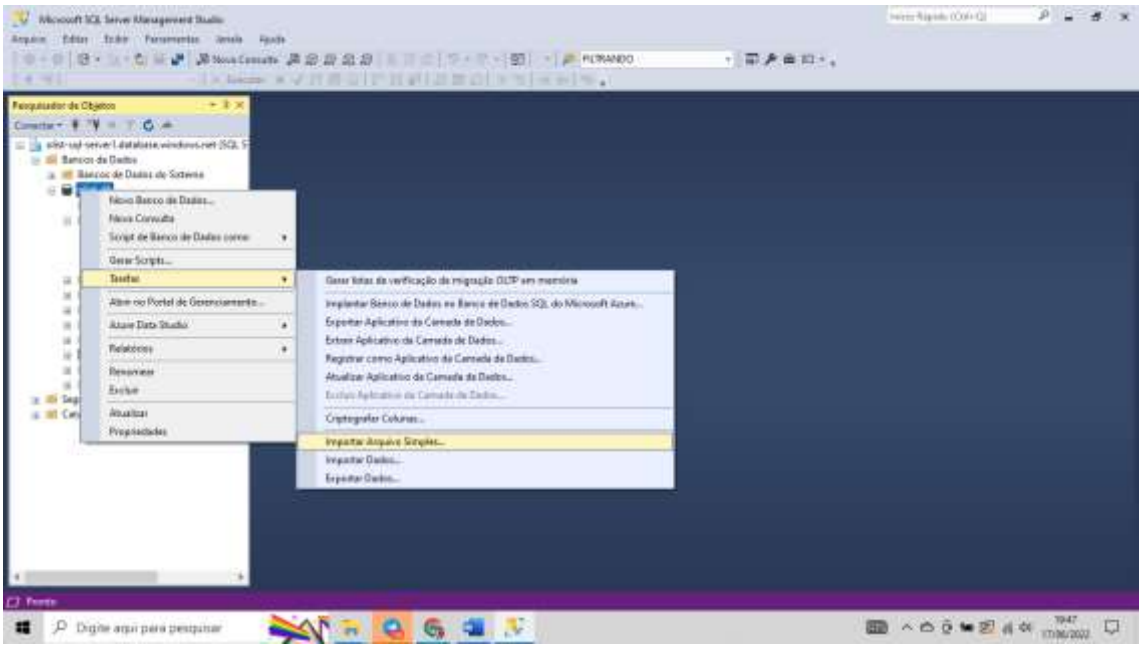


Obrigatoriamente tem que entrar com a conta da Azure

AGORA ESTÁ CONECTADO



IMPORTANDO AS TABELAS

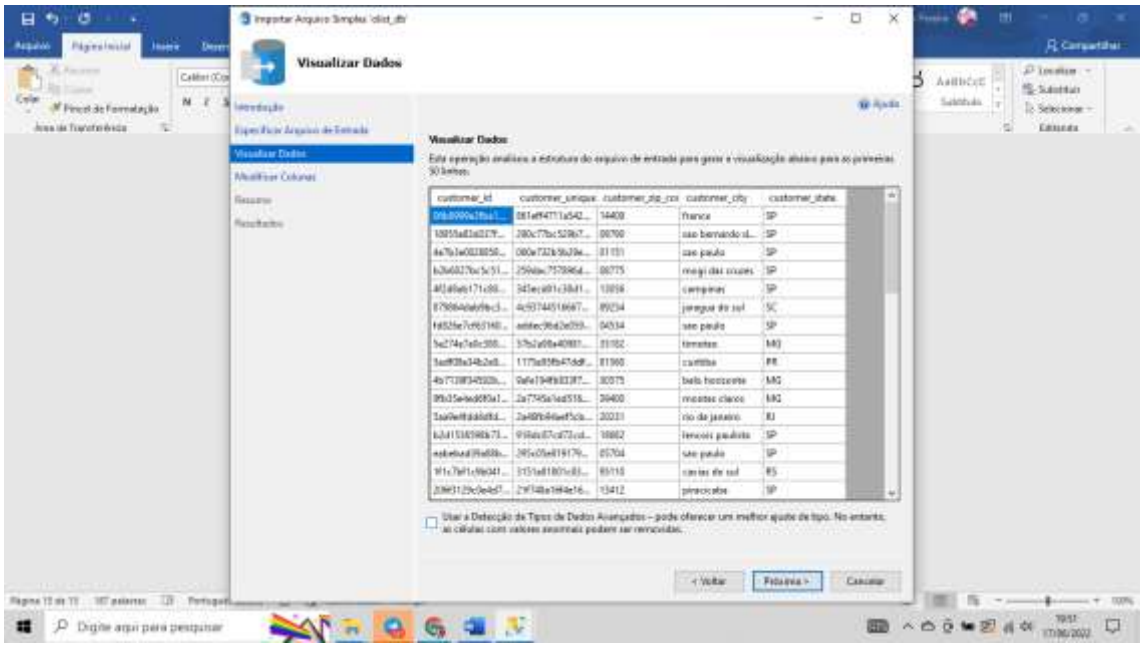


CLICAR EM AVANÇAR

SELECIONAR O ARQUIVO (TABELA)

REMOVER A SELEÇÃO DE DETECÇÃO DE TIPOS DE DADOS PARA NÃO TER ERRO NA IMPORTAÇÃO

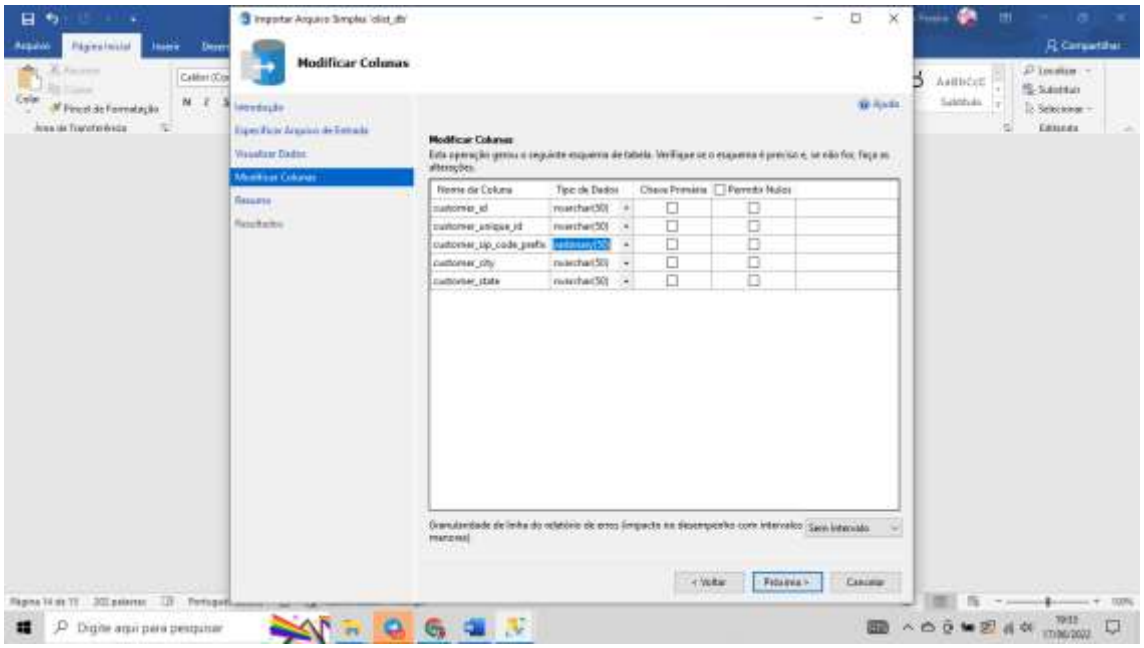
CLICA EM PROXIMO



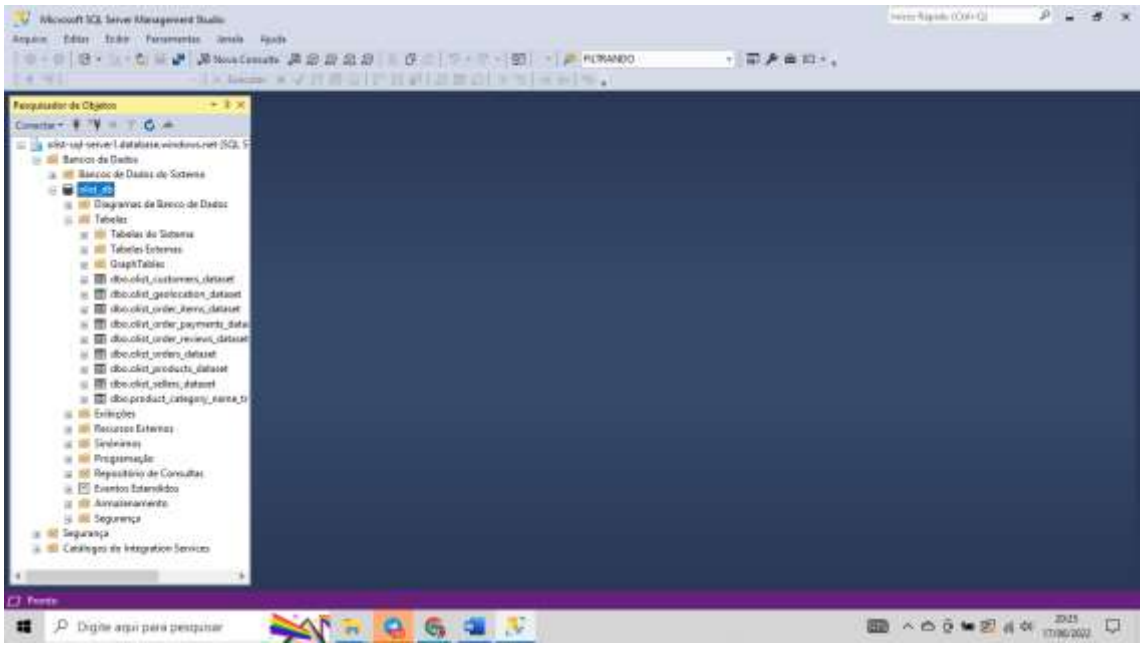
SELECIONAR TODOS OS CAMPOS COMO NVARCHAR (50)

CLICAR EM PROXIMO E FAZER ESSE MESMO PROCESSO PARA TODAS AS TABELAS

EM CASO DE ERRO NO CARREGAMENTO SELECIONAR OS NULOS



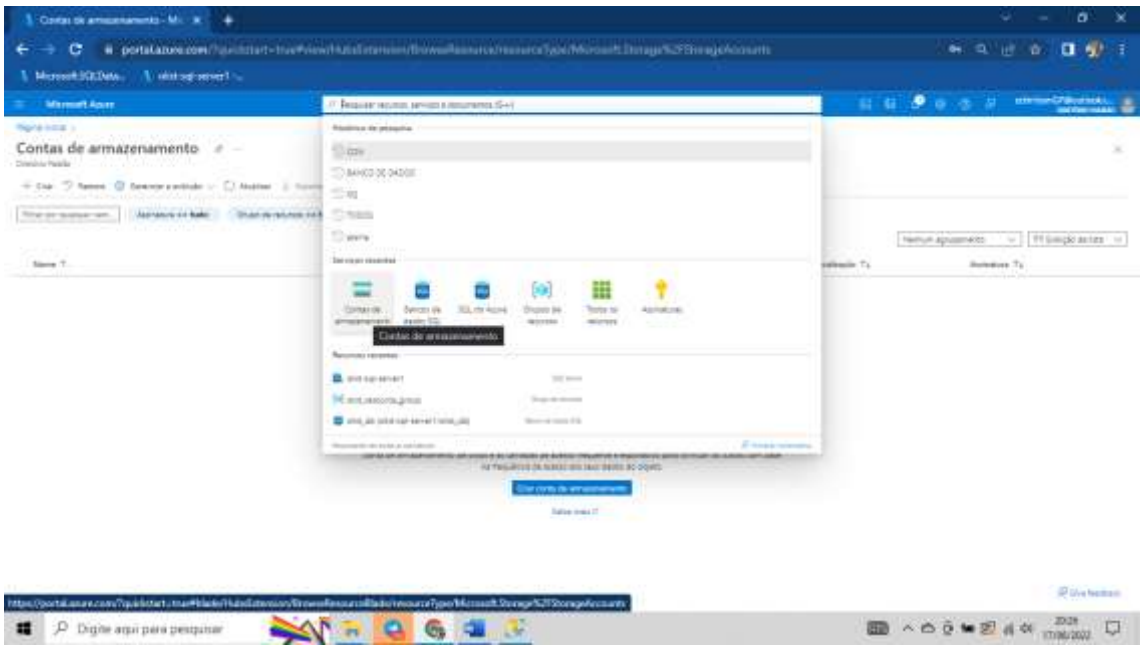
TODAS AS TABELAS CARREGADAS



CRIAR O DATA LAKE

PESQUISAR POR CONTAS DE ARMAZENAMENTO

E CLICAR EM CRIAR



Nome da conta de armazenamento
oliststorageacount2

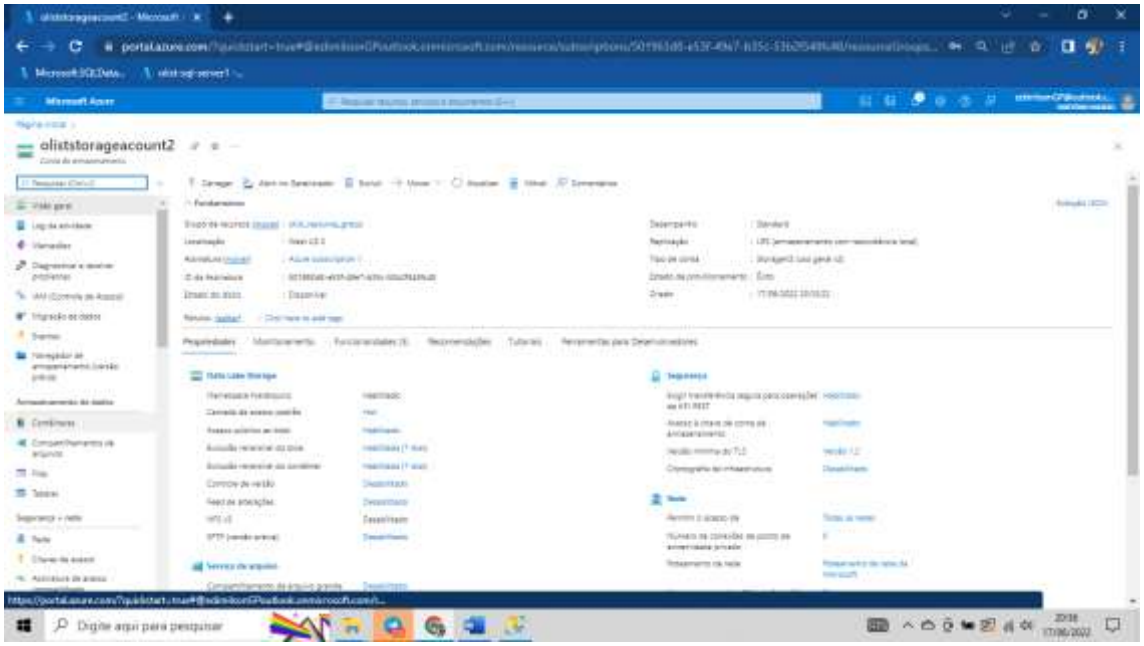
clicar em avançar

selecionar Data Lake Storage Gen2

Agora em oliststorageacount2

Clicar em Conteineres

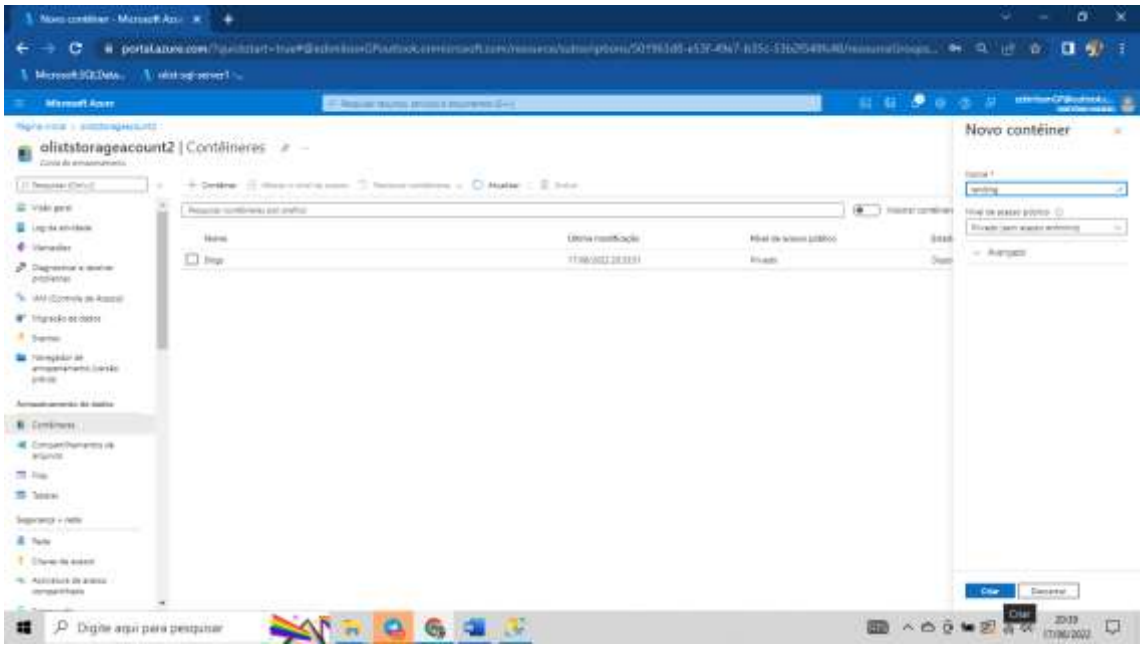
Para criar Landing, Processing, Acuretad



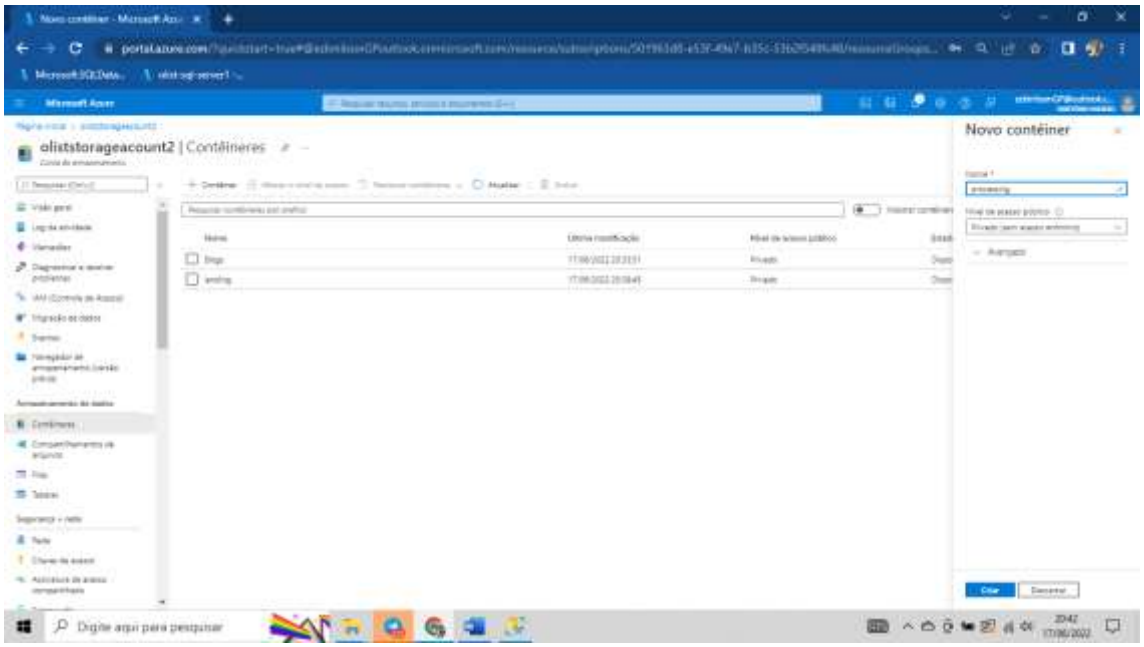
Clicar em +Conteiner e colocar o nome nesse caso é a

LANDING

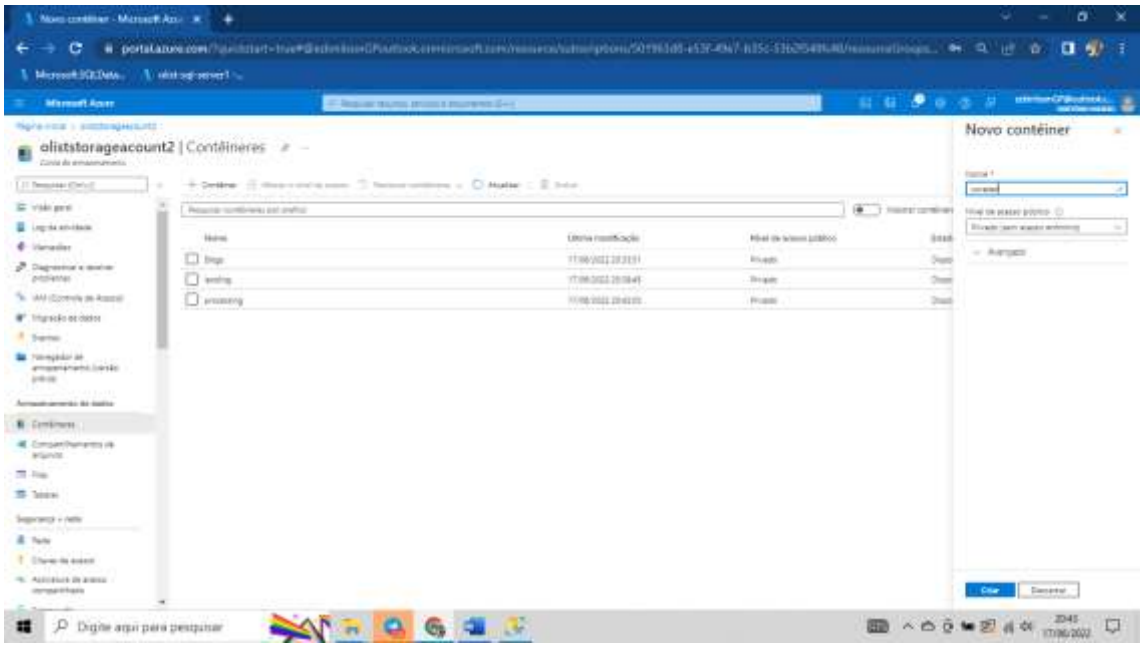
E criar



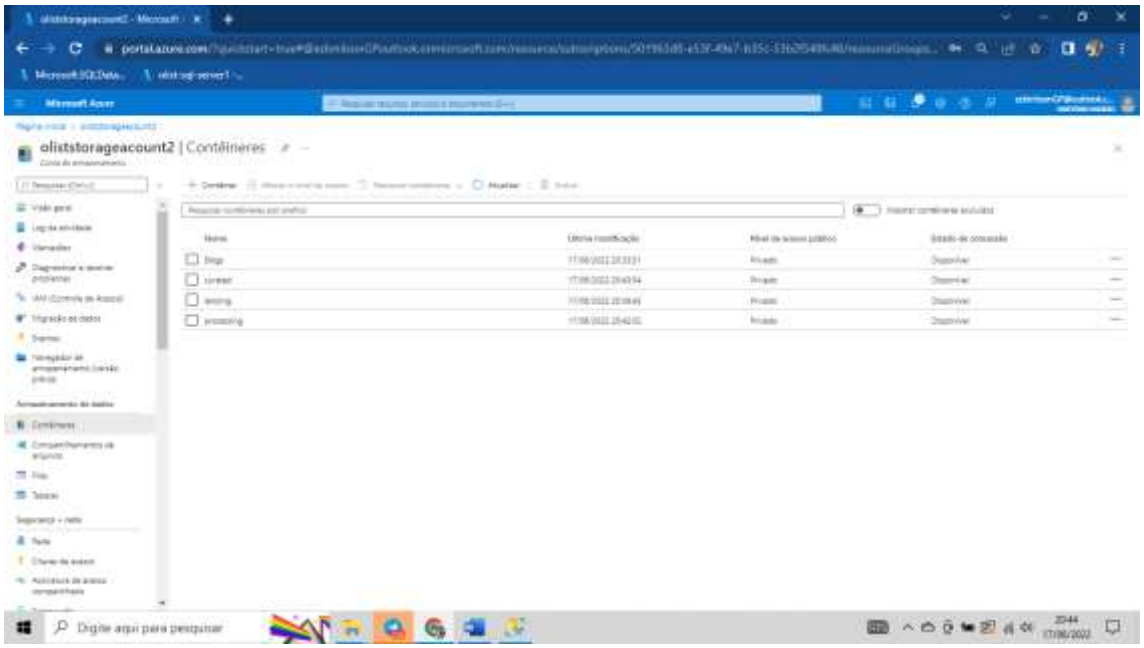
PROCESSING



CURATED

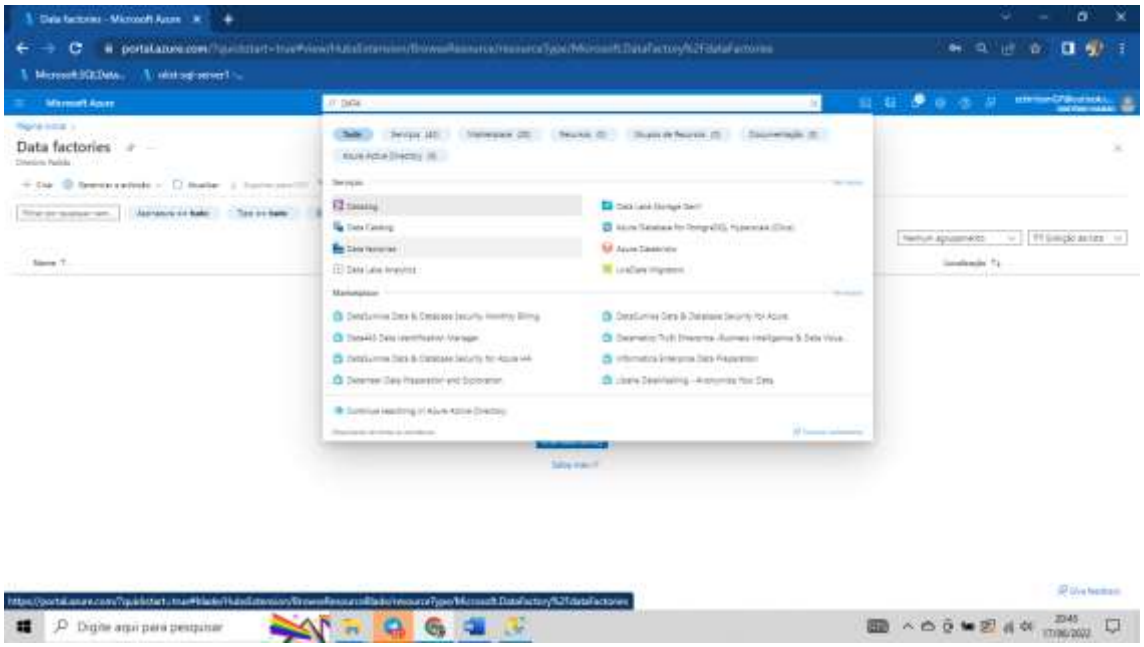


CONTEINÊRES CRIADOS

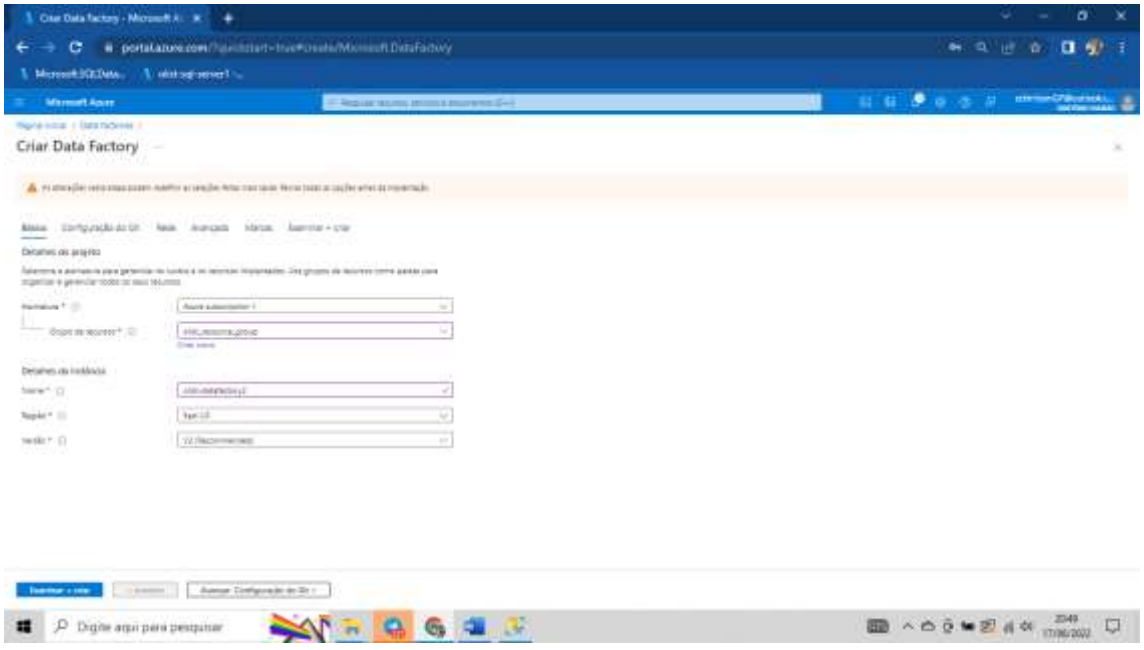


DATAFACTORY

É UMA FERRAMENTA DE ETL ELE AJUDA A TRANSFERIR OS DADOS DE UM LUGAR PARA OUTRO DE DIVERSAS FONTES DISPONIVEIS

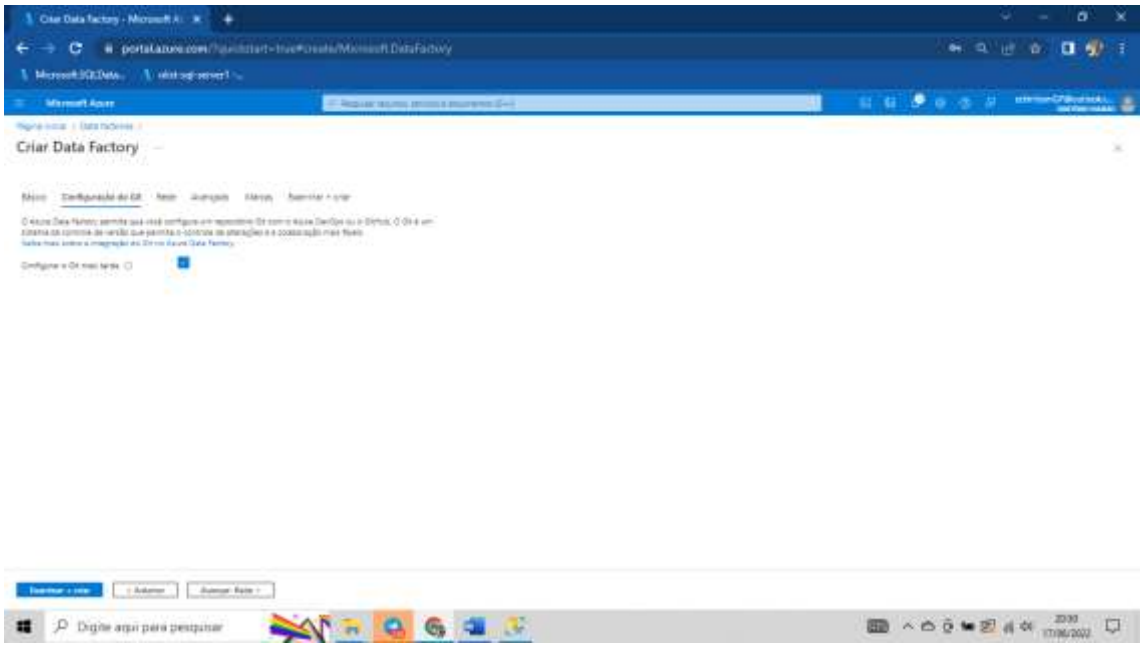


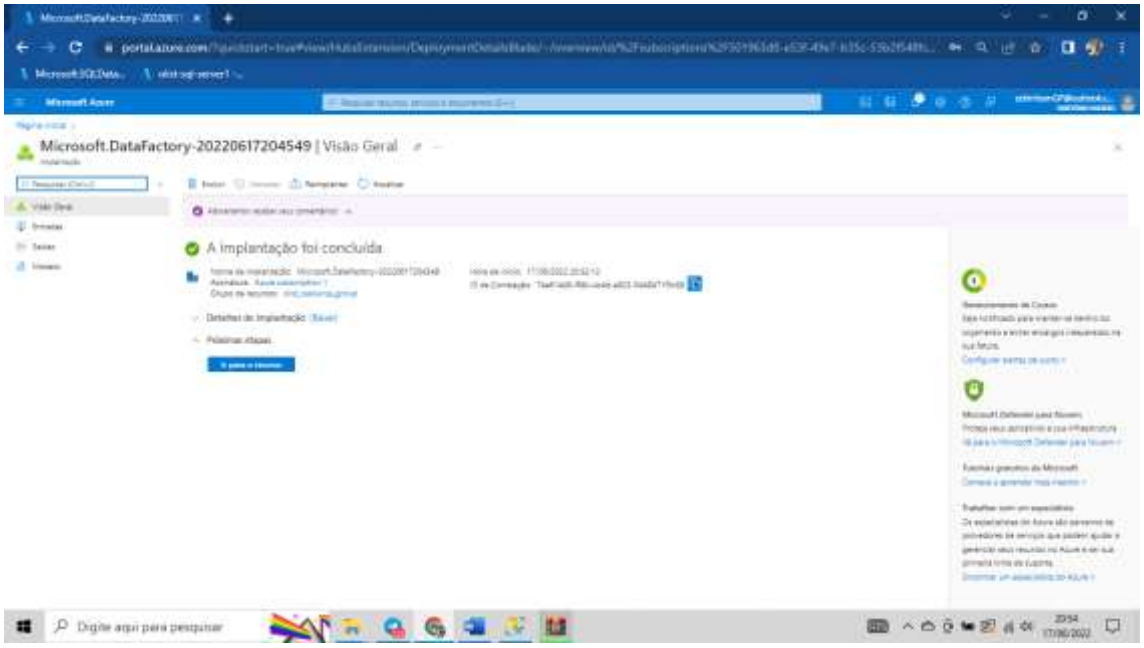
olist-datafactory2



Avançar e selecionar configurar o git mais tarde

E criar





ABRIR O DATA FACTORY – para fazer o ETL

E CLICAR EM ABRIR O ESTUDIO DO AZURE DATA FACTORY

Clicar em NEW PIPELINE

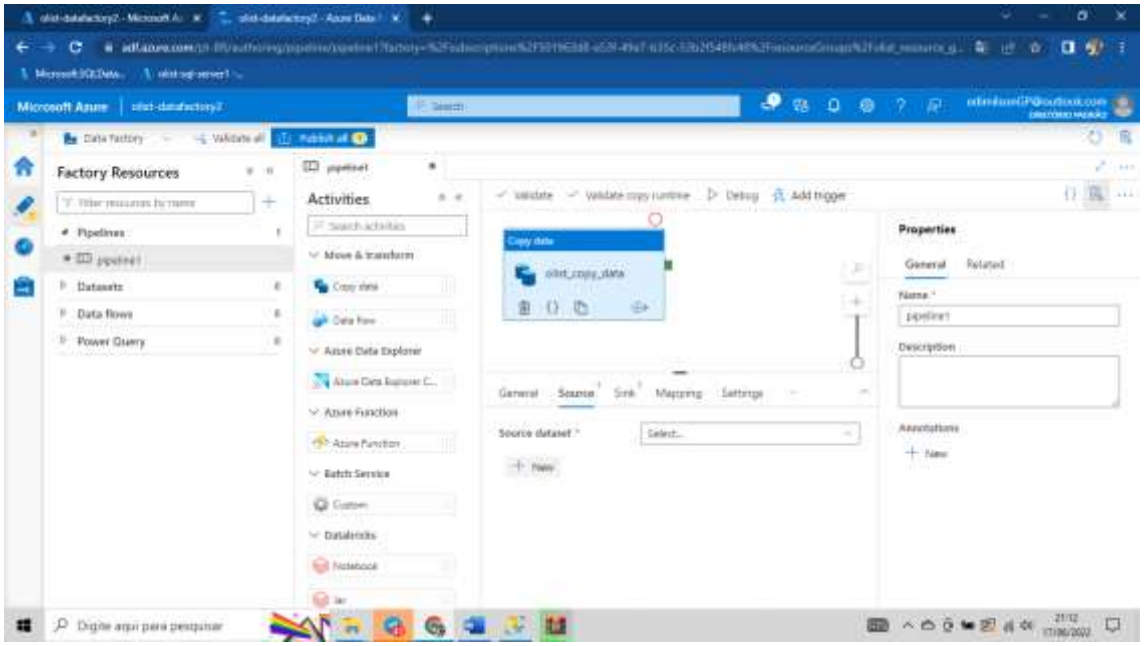
COPIANDO OS DADOS

Clicar e arrastar o ícone para a área e configurar

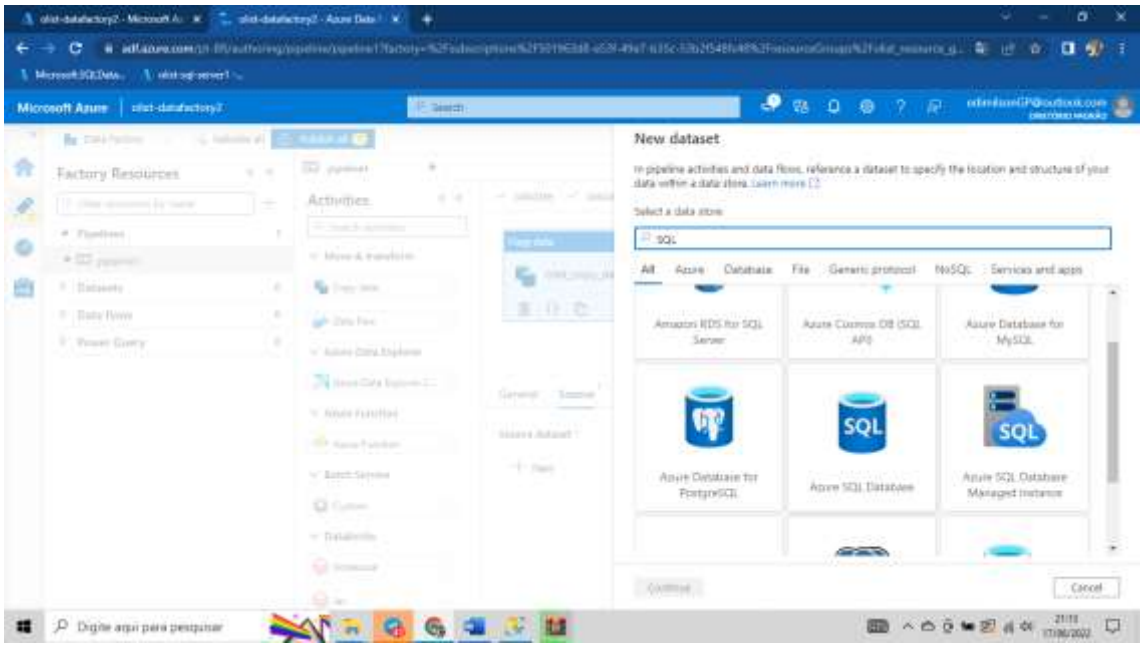
olist_copy_data

SELECIOR A ORIGEM DOS DADOS EM SOURCE

NEW

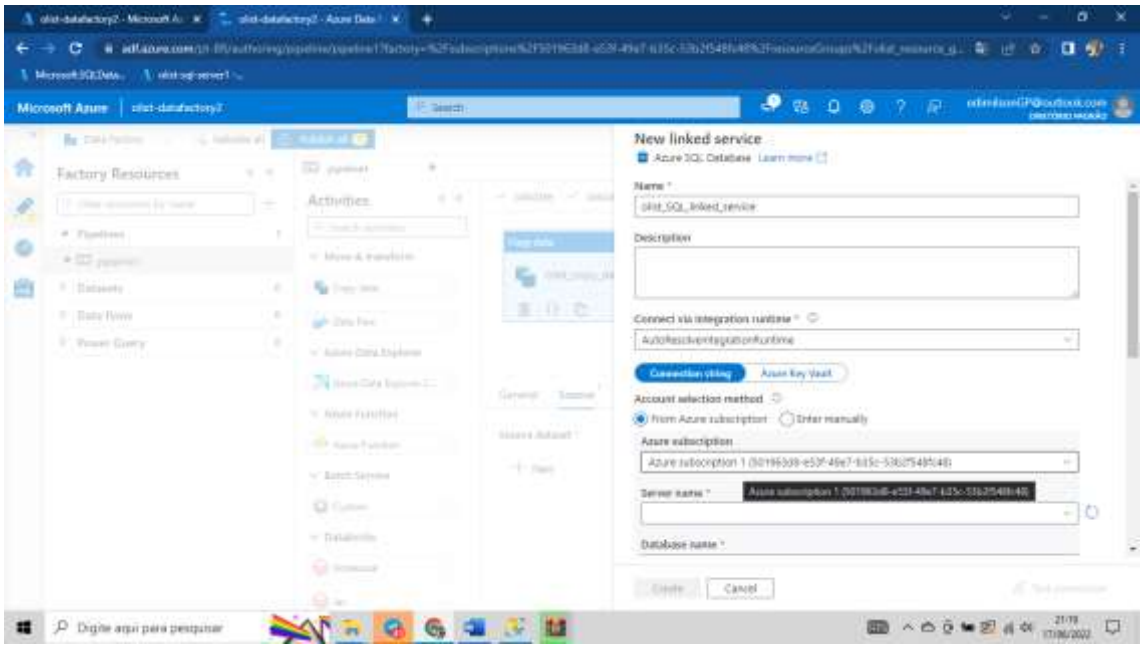


Azure SQL

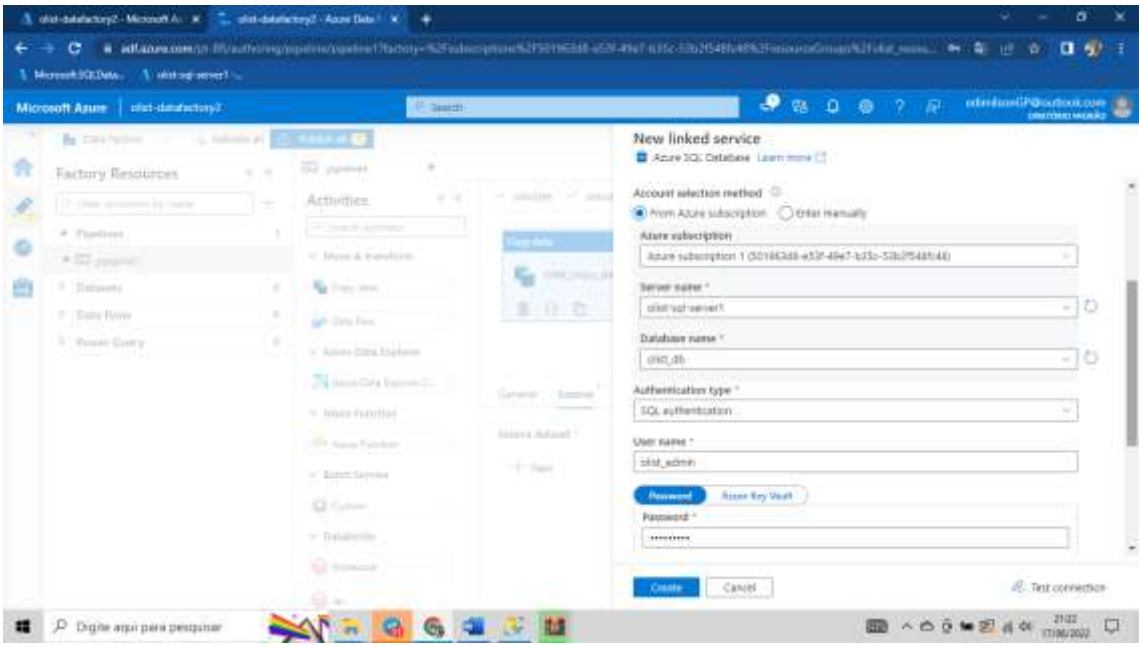


olist_SQL_linked_service1

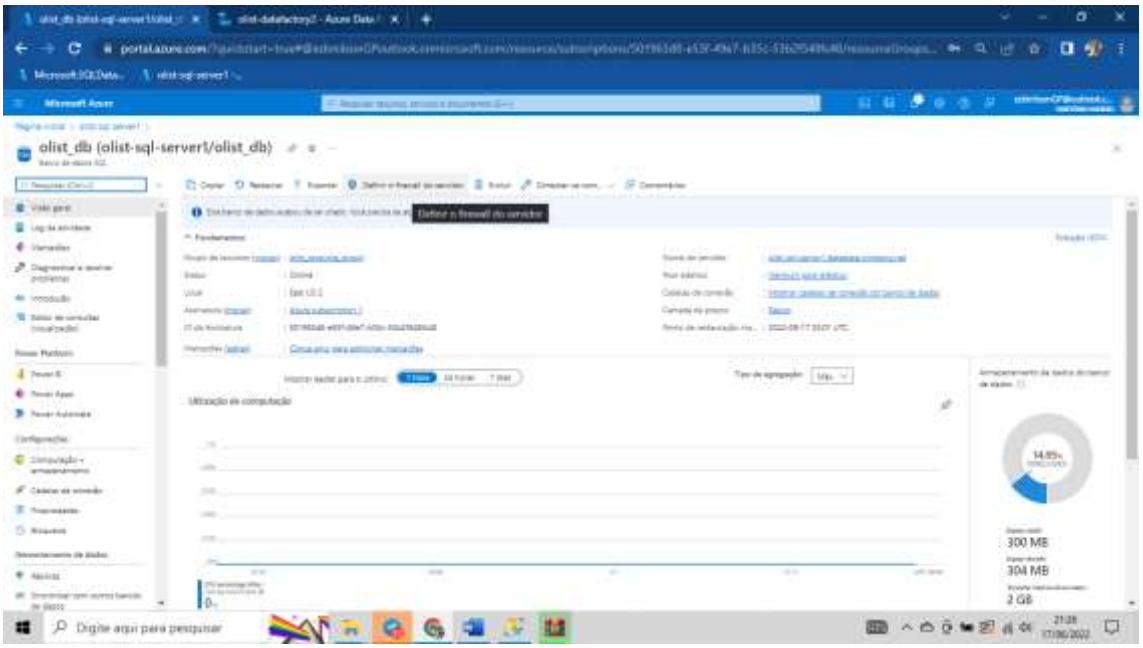
CONEXÃO ASSINATURA DO AZURE



Clicar em testar a conexão



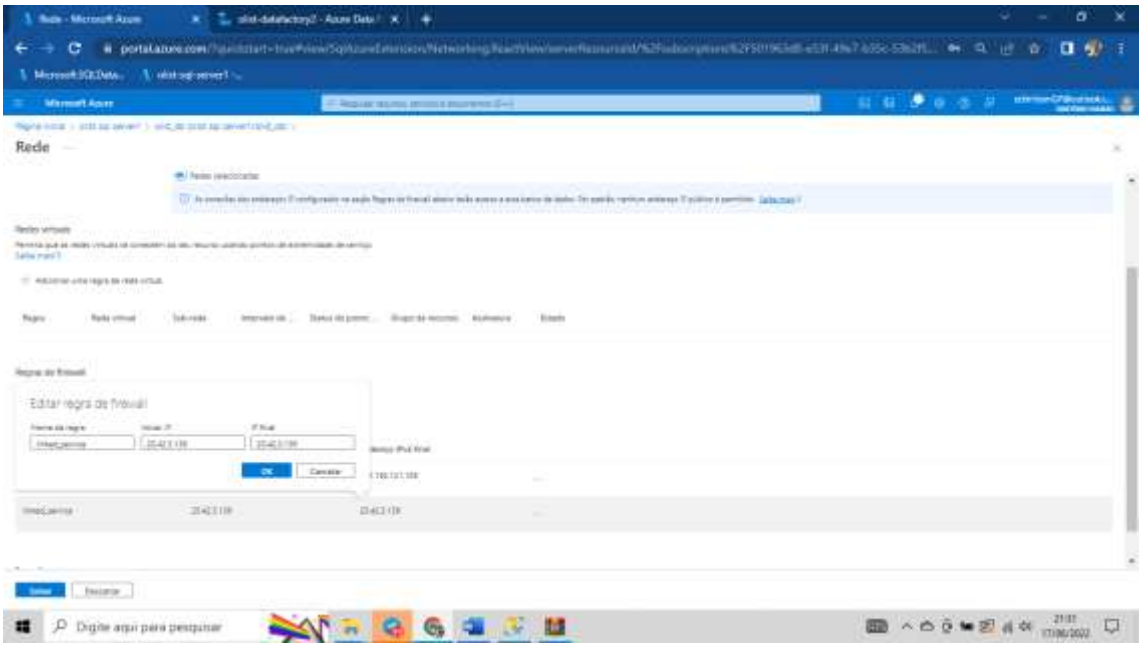
ACESSANDO PERMISSÃO E REGRA DE FIREWALL



ADICIONAR REGRA

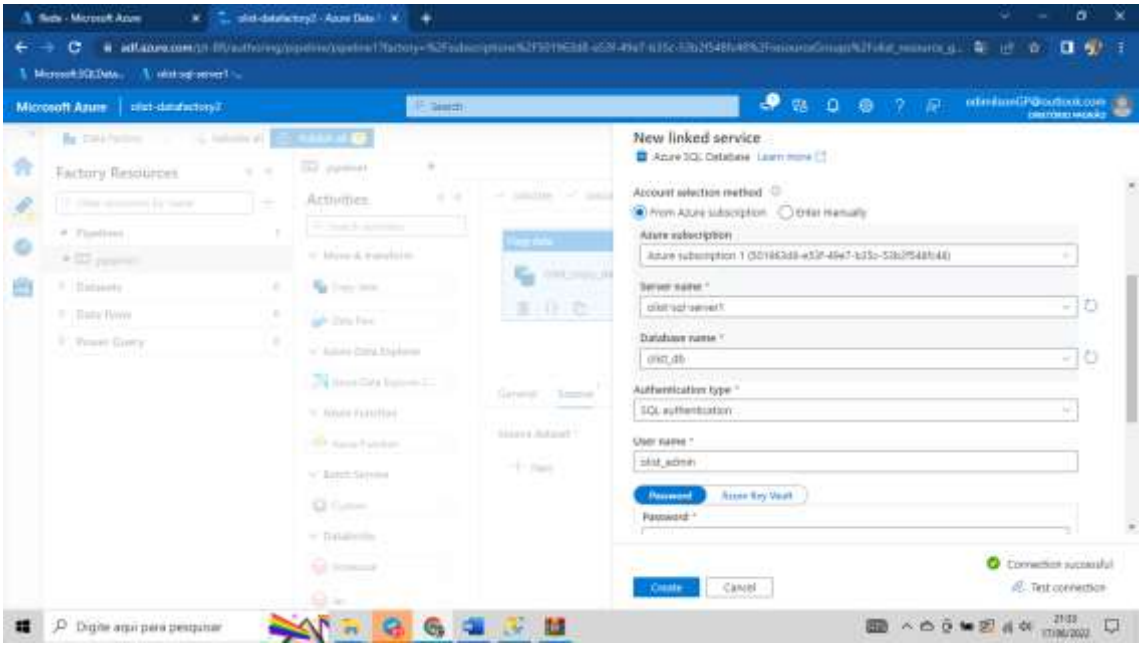
Colocando o IP que aparece lá no erro do teste de conexão '20.42.3.136'

E clicar em salvar

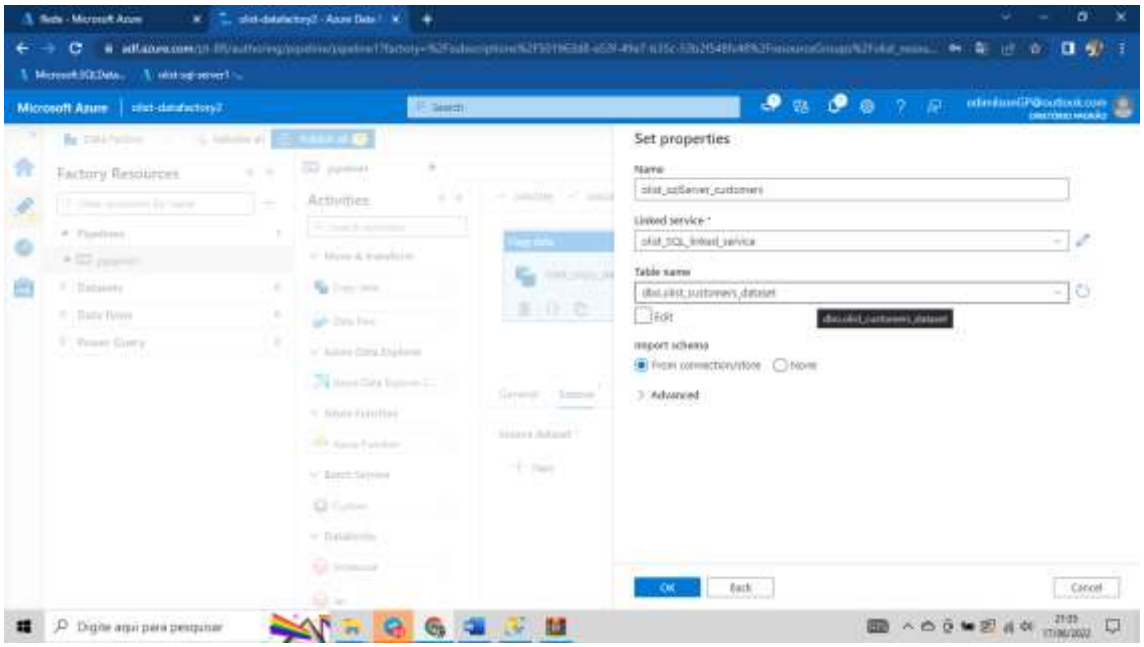


Testando novamente a conexão

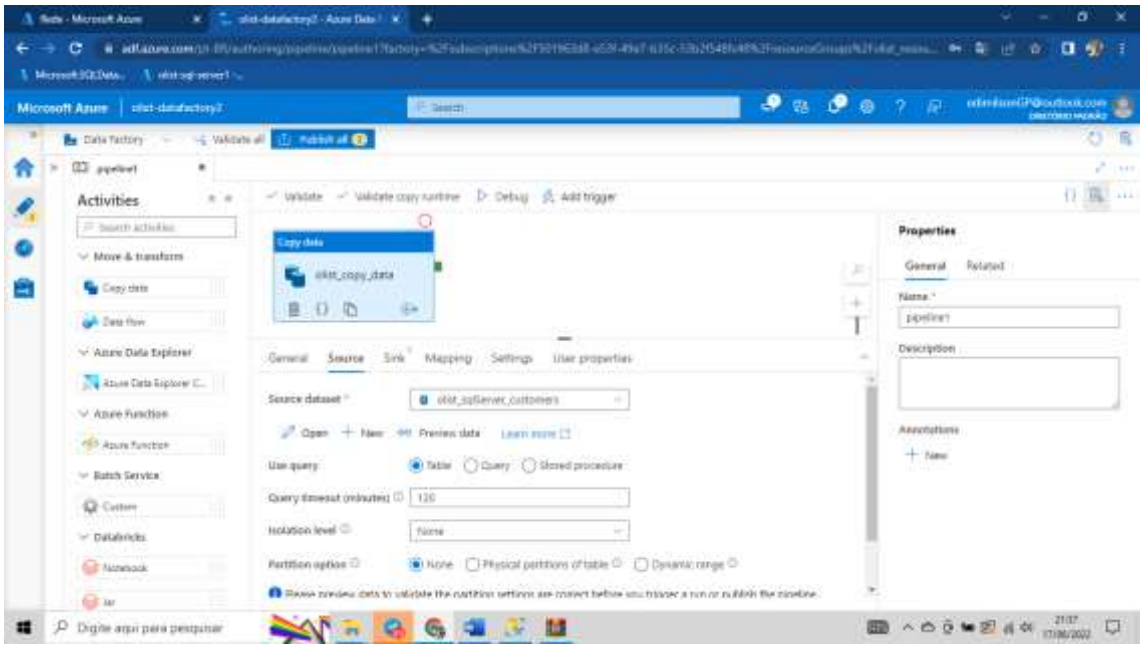
Conectado com sucesso e clicar em Criar



Agora eu escolho a tabela que eu quero



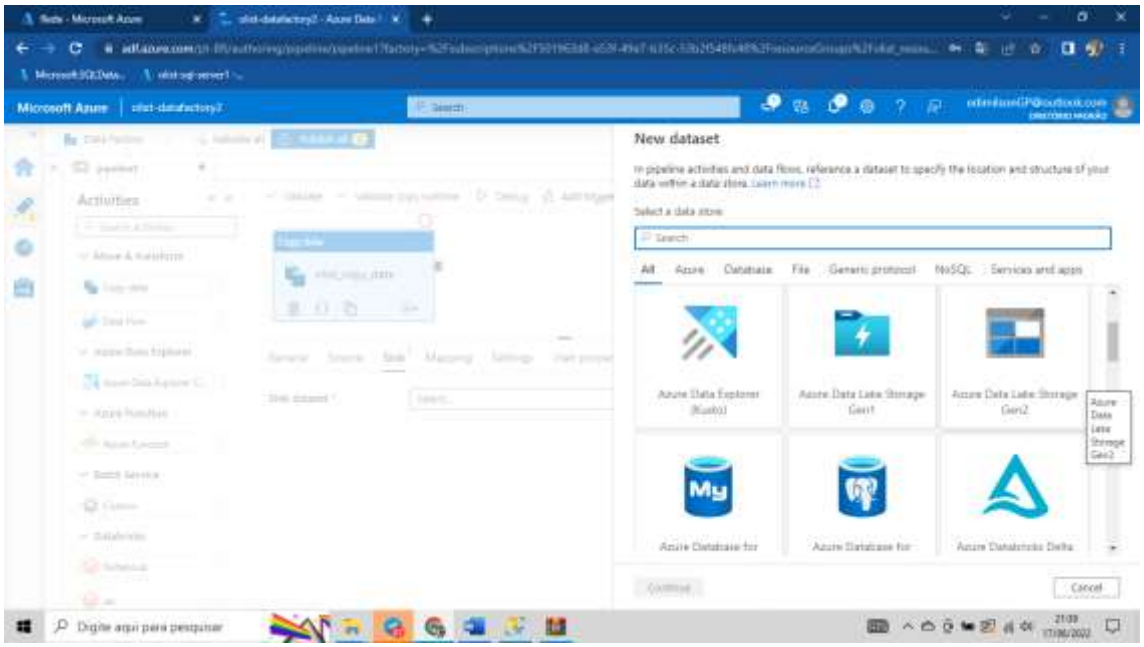
Origem criada



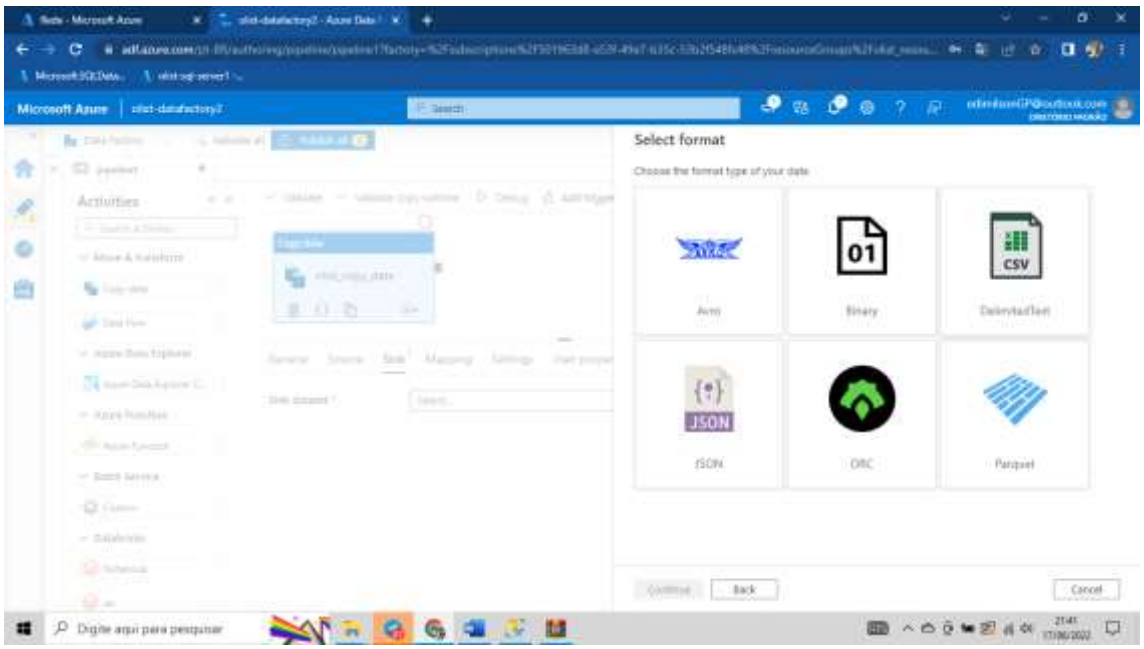
Agora para onde vou enviar os dados clicar em SINK

Clicar em NEW e

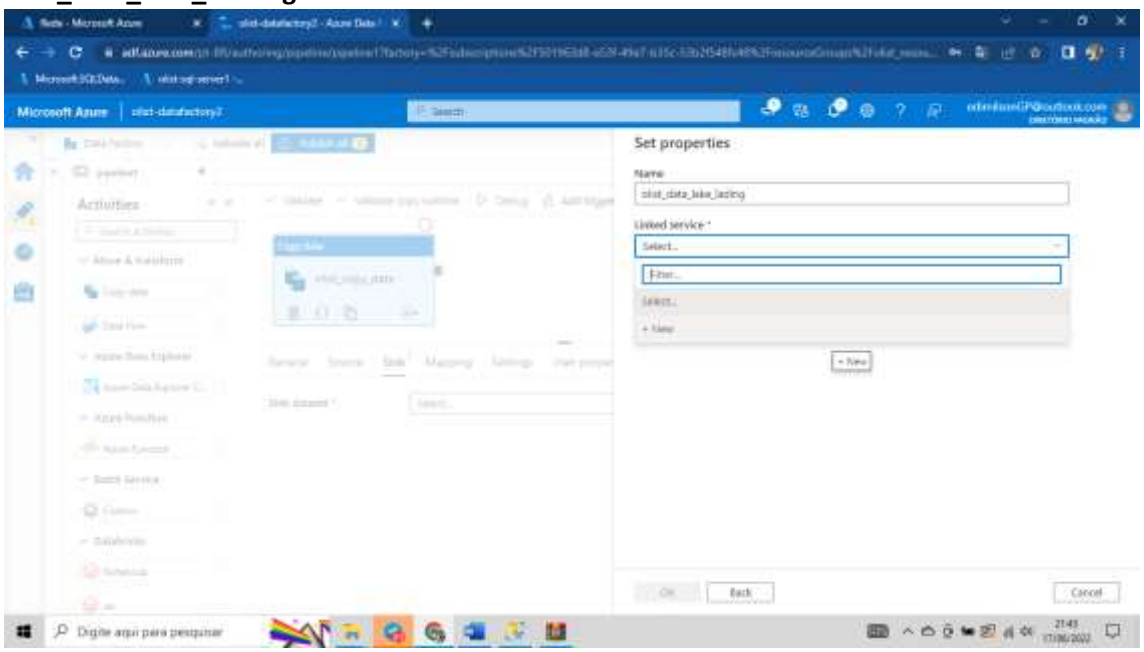
Pesquisar pelo Azure Data Lake Storage Gen2



Selecionar o tipo de arquivo que estamos trabalhando no caso as tabelas estão no formato CSV



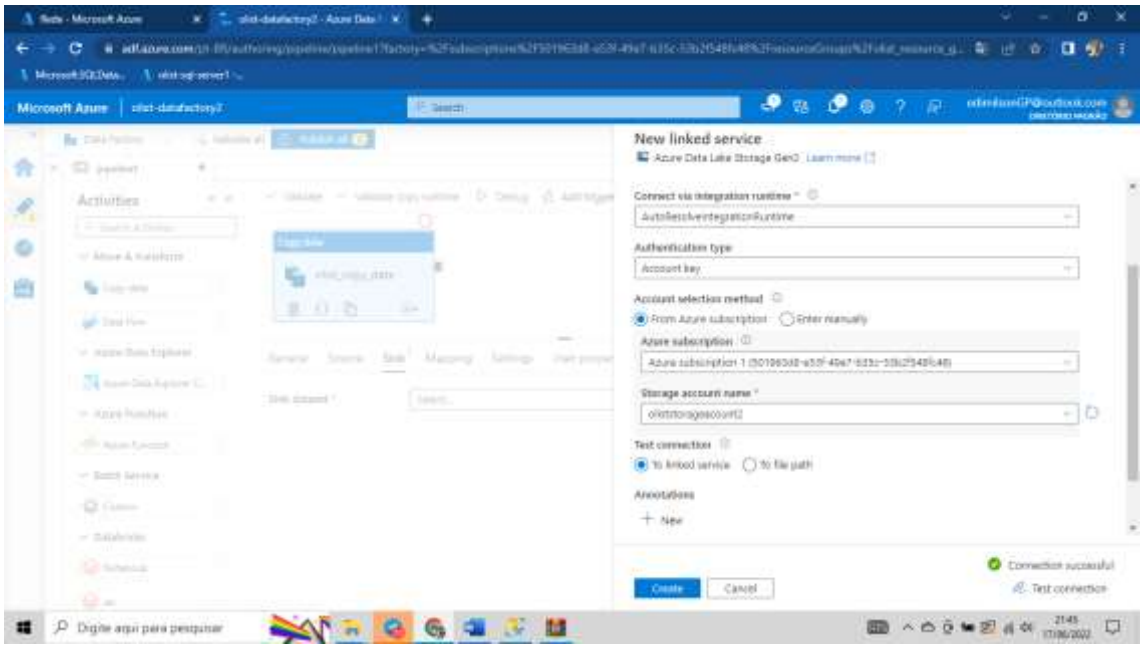
olist_data_lake_landing1

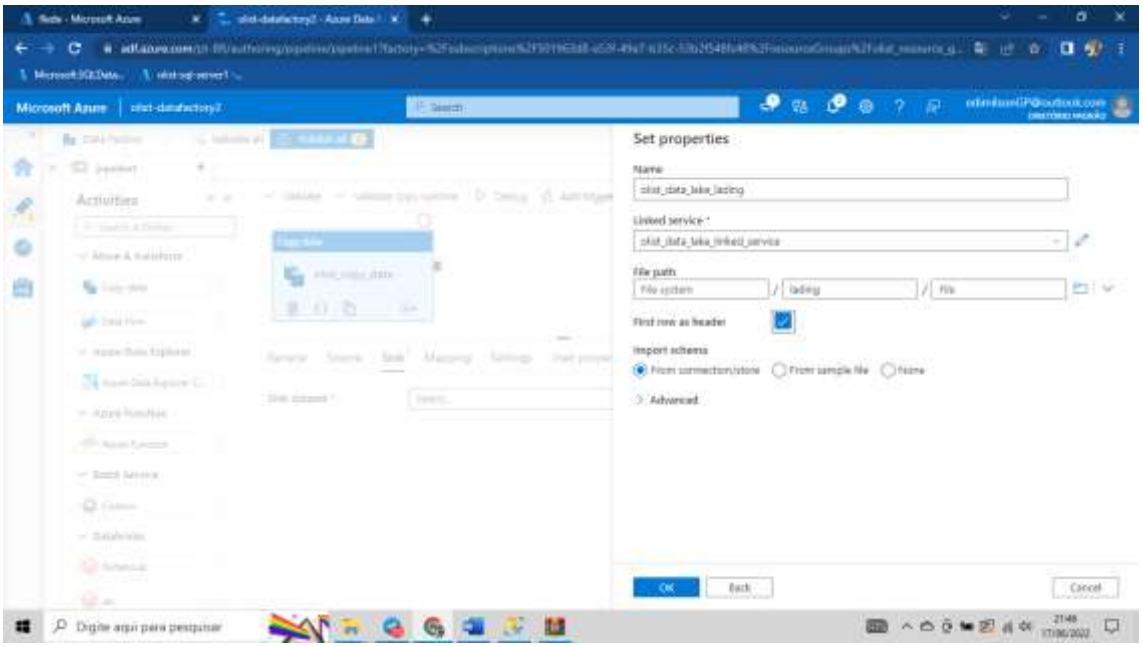


olist_data_lake_linked_service1

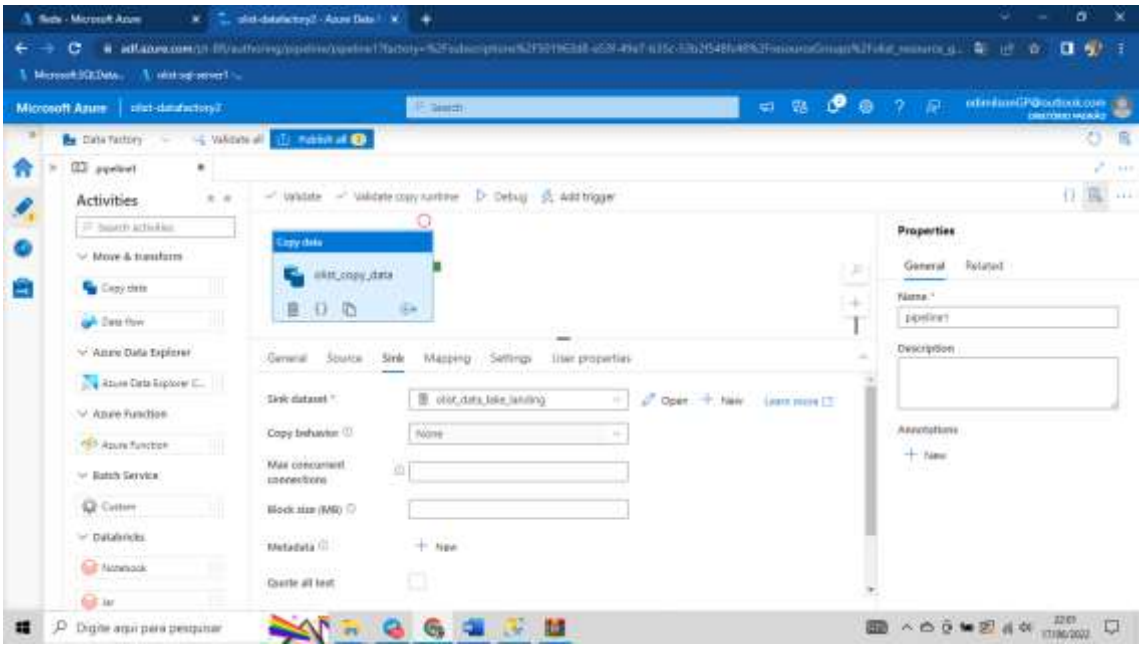
selecionando as informações

testando a conexão e criar

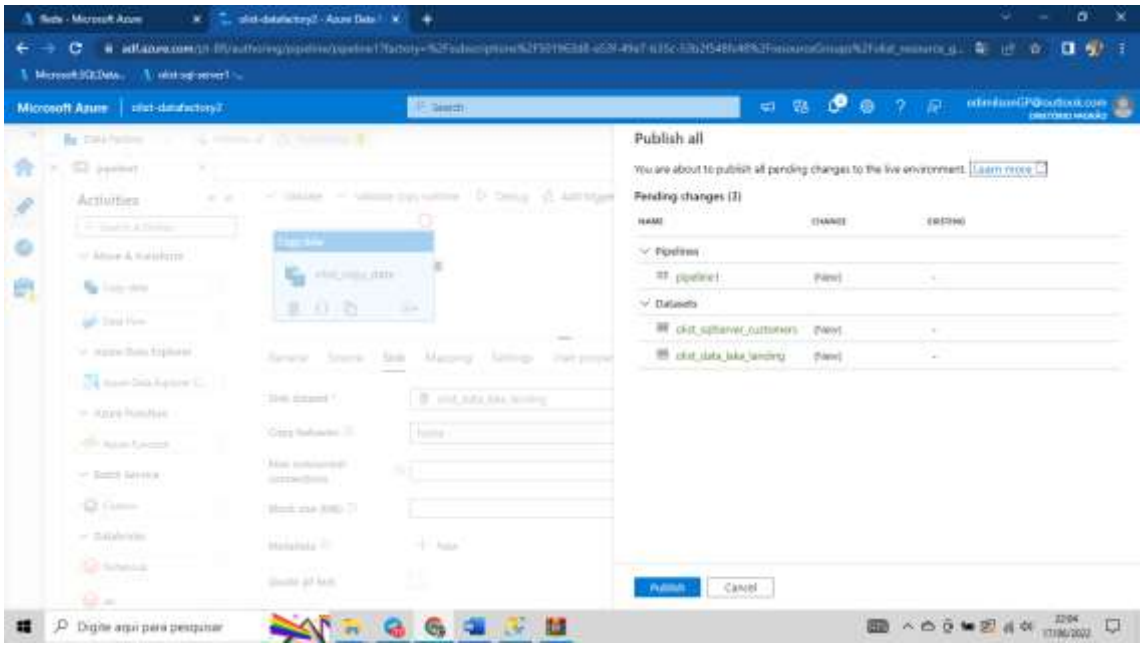




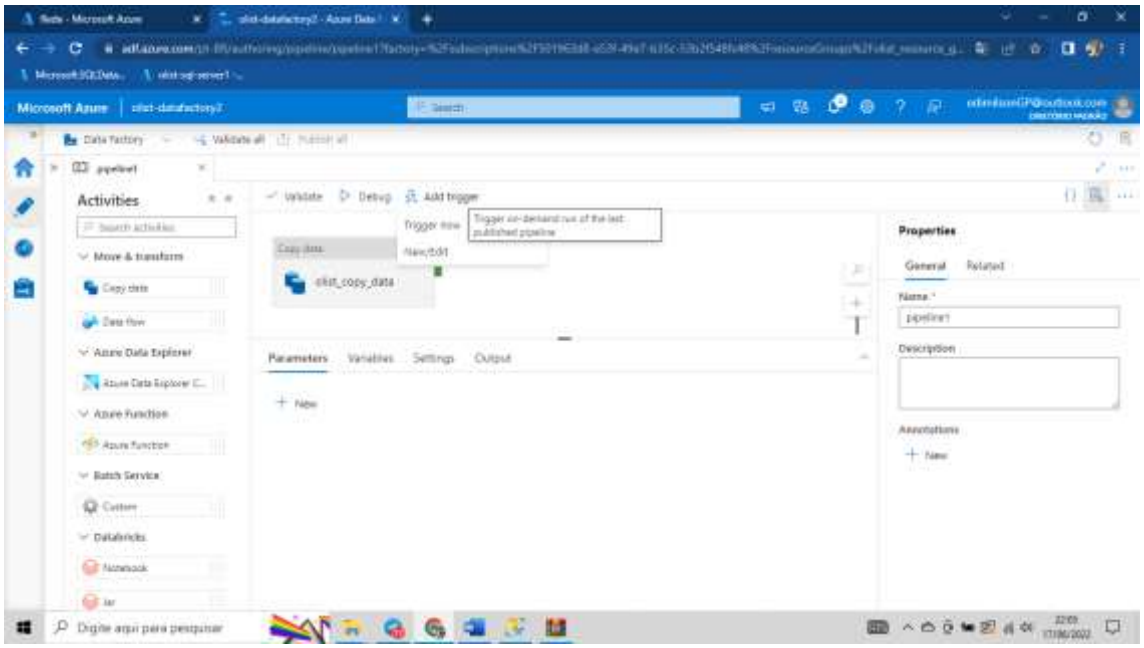
CONFIGURAÇÃO DE DESTINO EFETUADA



PARA TESTARMOS O PIPELINE TEMOS QUE PUBLICAR TODAS AS ATUALIZAÇÕES

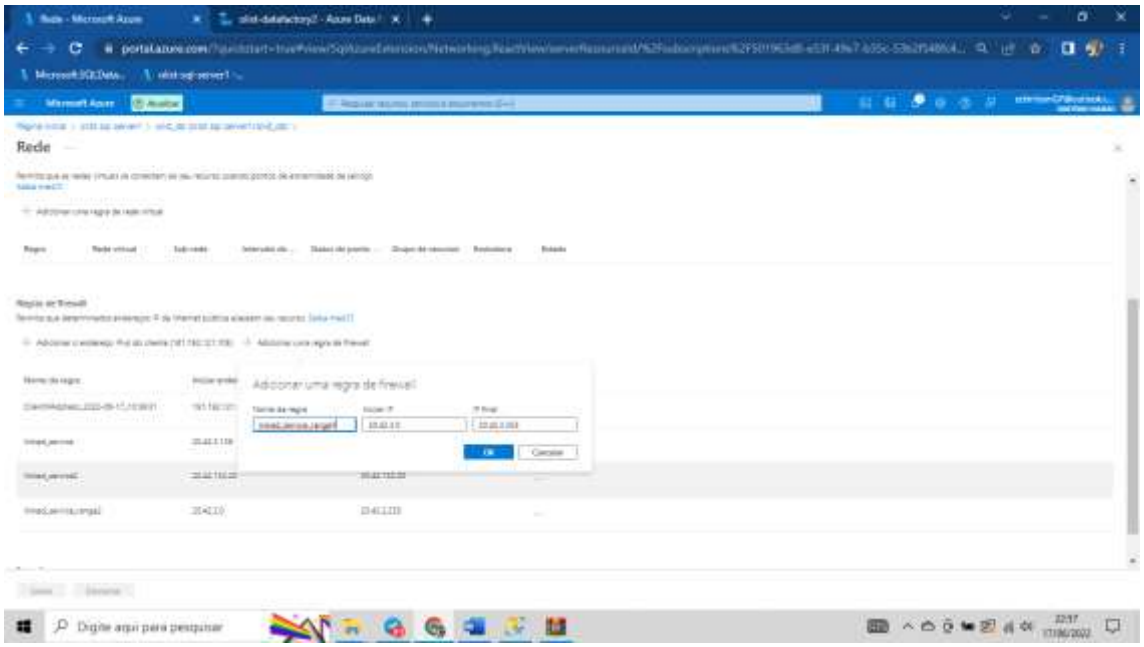


PARA EXECUTAR VAMOS ADICIONAR O GATILHO ADD TRIGGER NOW



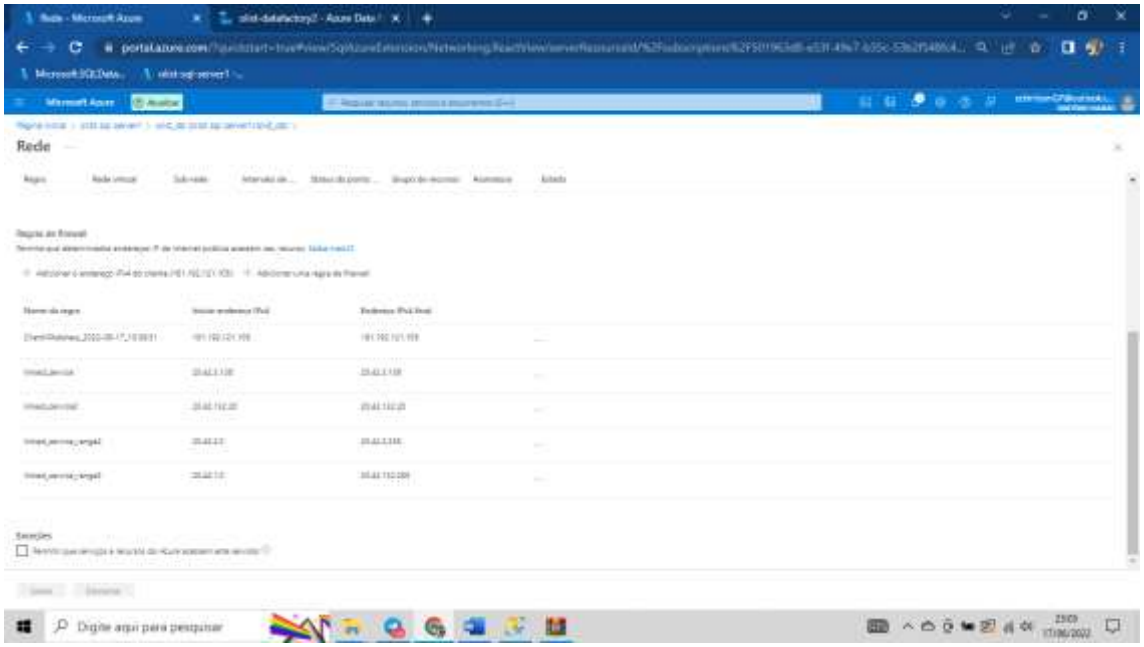
CONTINUA APRESENTANDO ERRO DE IP´S

ENTÃO TEREMOS QUE LIBERAR POR FAIXA



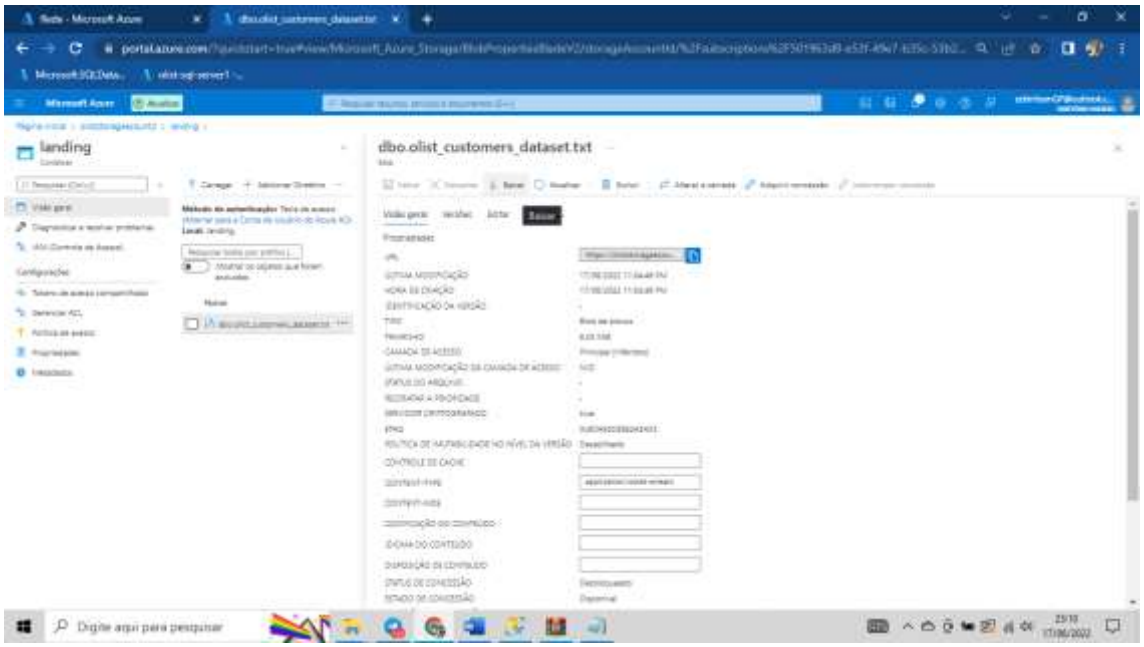
CONEXÃO FUNCIONANDO

COM AS FAIXAS INFORMADAS A BAIXO



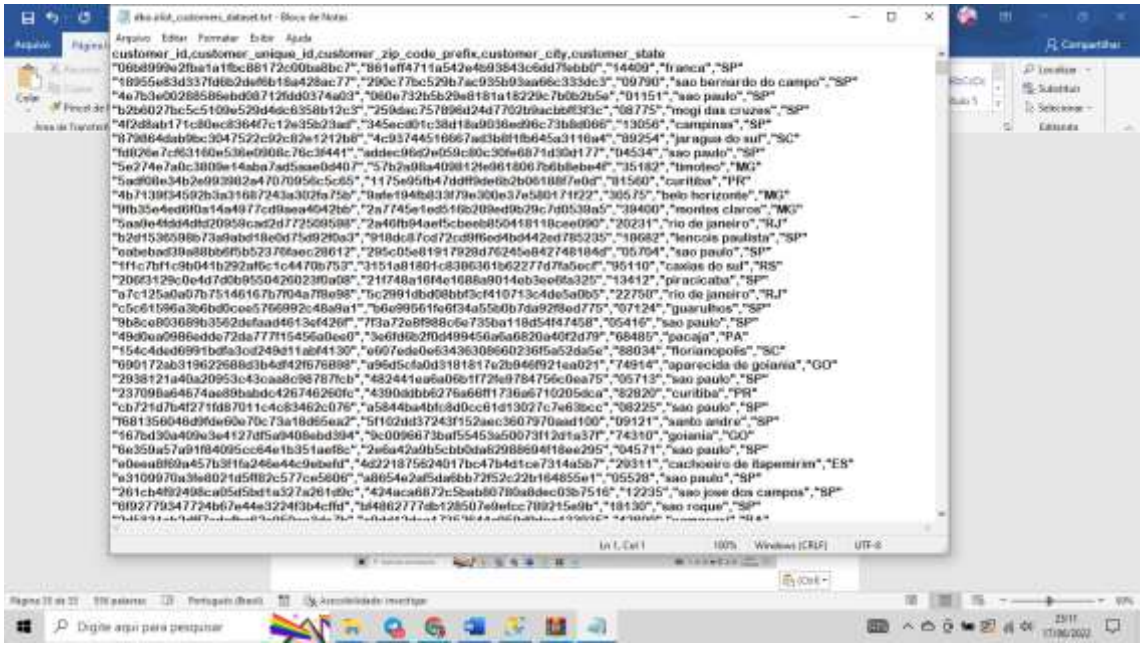
PARA VERIFICAR SE OCORREU TUDO CERTO COM O ETL

IR PARA PAGINA INICIAL EM OLISTSTORAGEACOUNT2 EM CONTAINERES E CLICAR NA LANDING E BAIXAR UMA COPIA



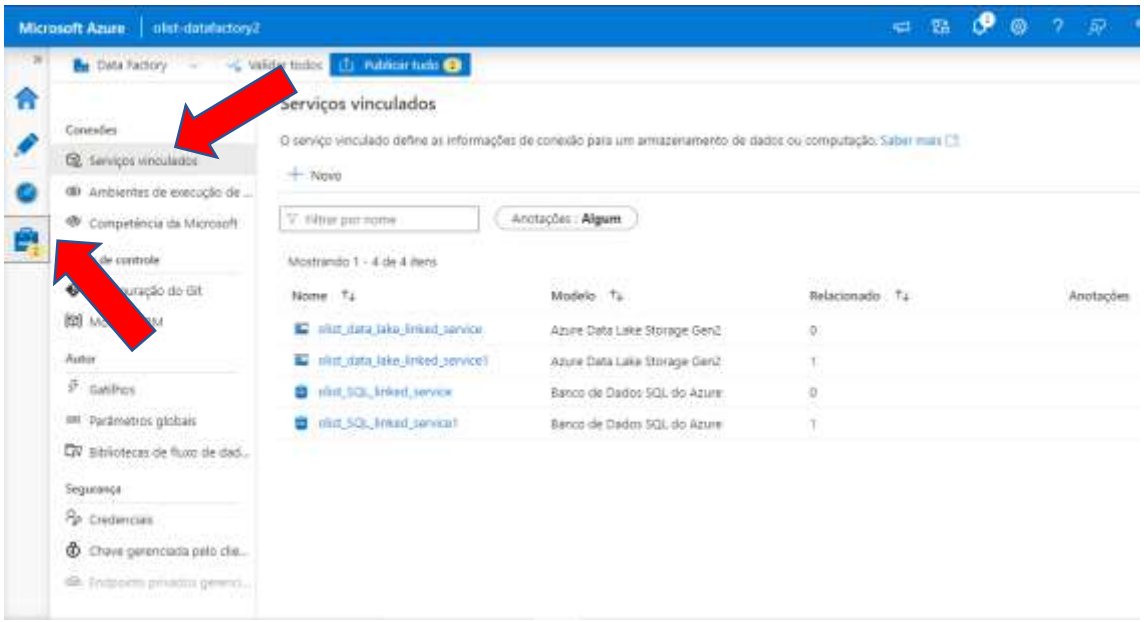
ABRINDO O ARQUIVO PARA CONFERIR

ENTÃO ESTÁ FUNCIONANDO

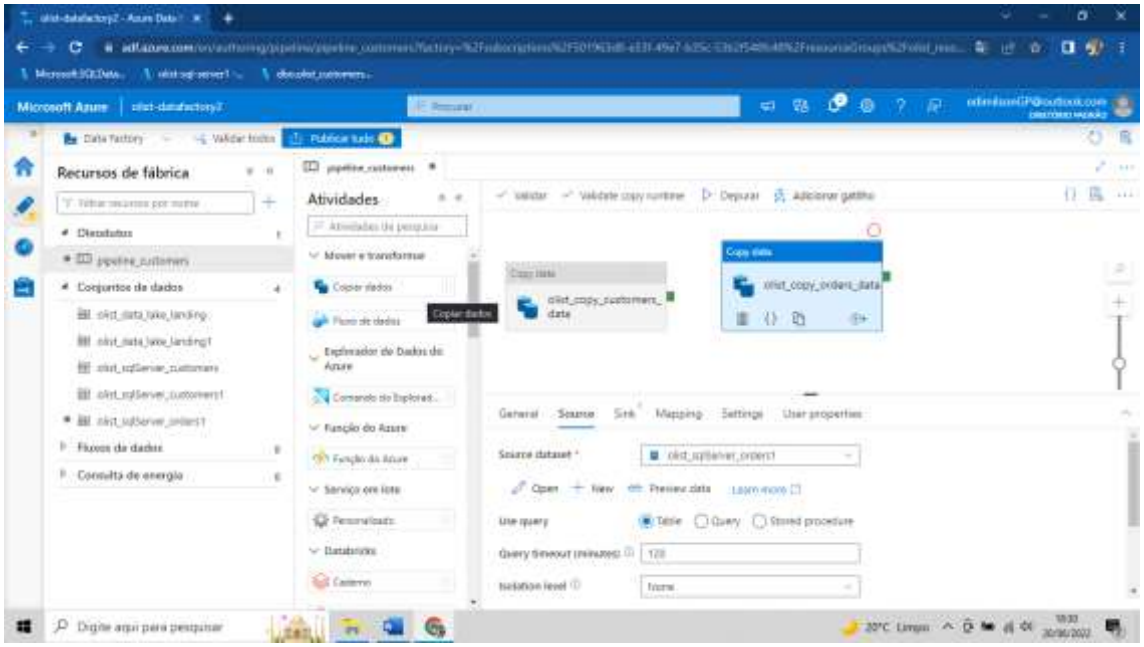


AGORA VAMOS CARREGAR TODAS AS TABELAS DE UMA ÚNICA VEZ PARA GANHO DE TEMPO

CLICAR EM GERENCIAR E EM SERVIÇOS VINCULADOS

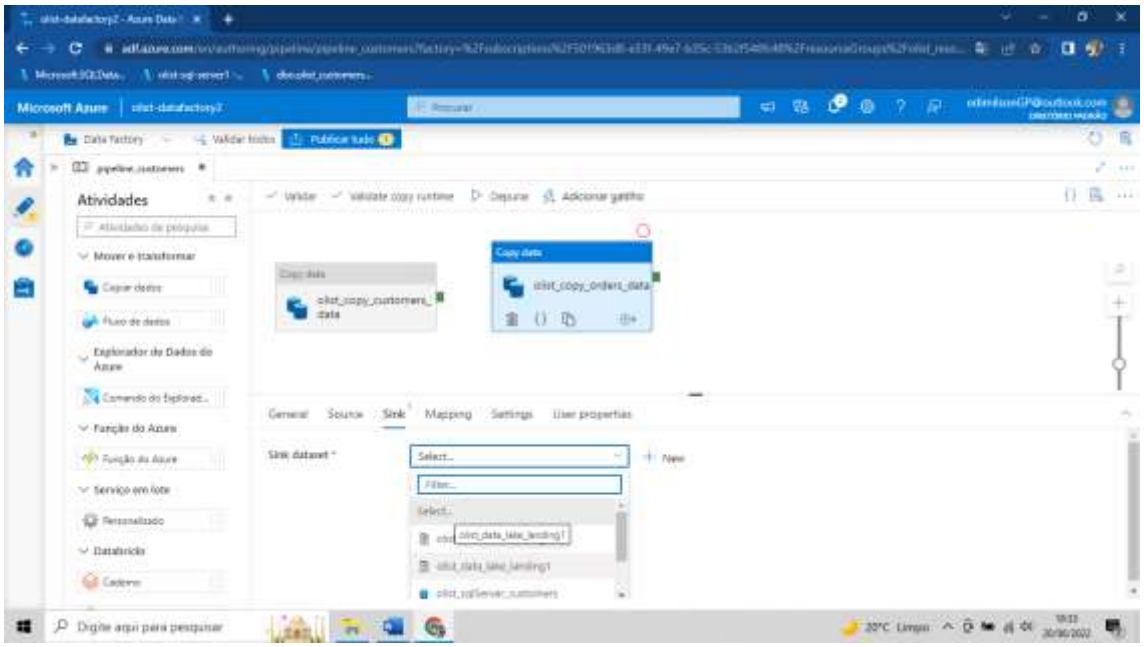


CRIAR UMA NOVA COPIA DE DADOS



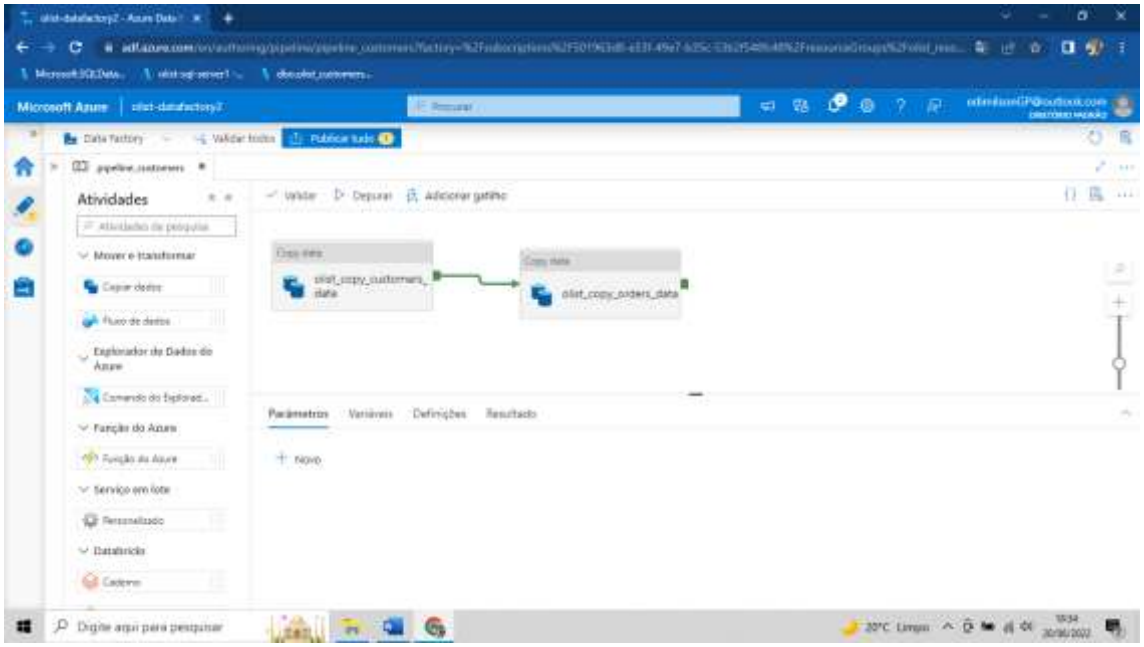
COMO JÁ EXISTE OS CAMPOS AGORA SÓ SELECIONAR OS CAMPOS CORRETOS

O DATALEKE DE LANDING JÁ EXISTE SÓ SELECIONAR NO DESTINO



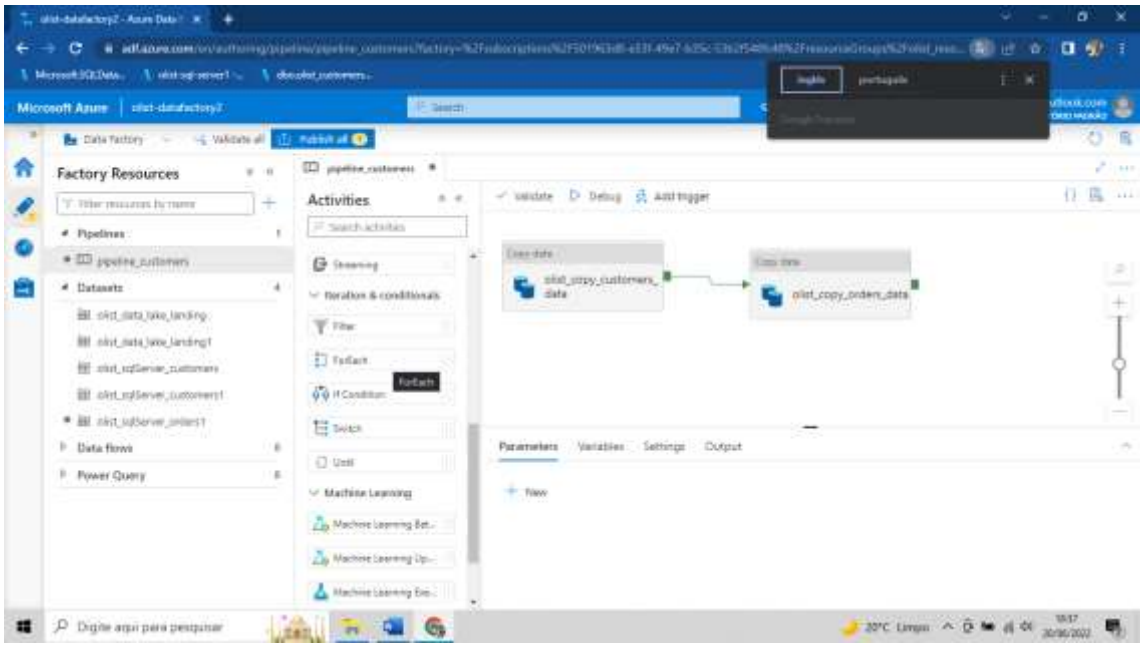
AGORA FAZENDO A CONEXÃO DOS DADOS

CLICANDO E ARRASTANDO



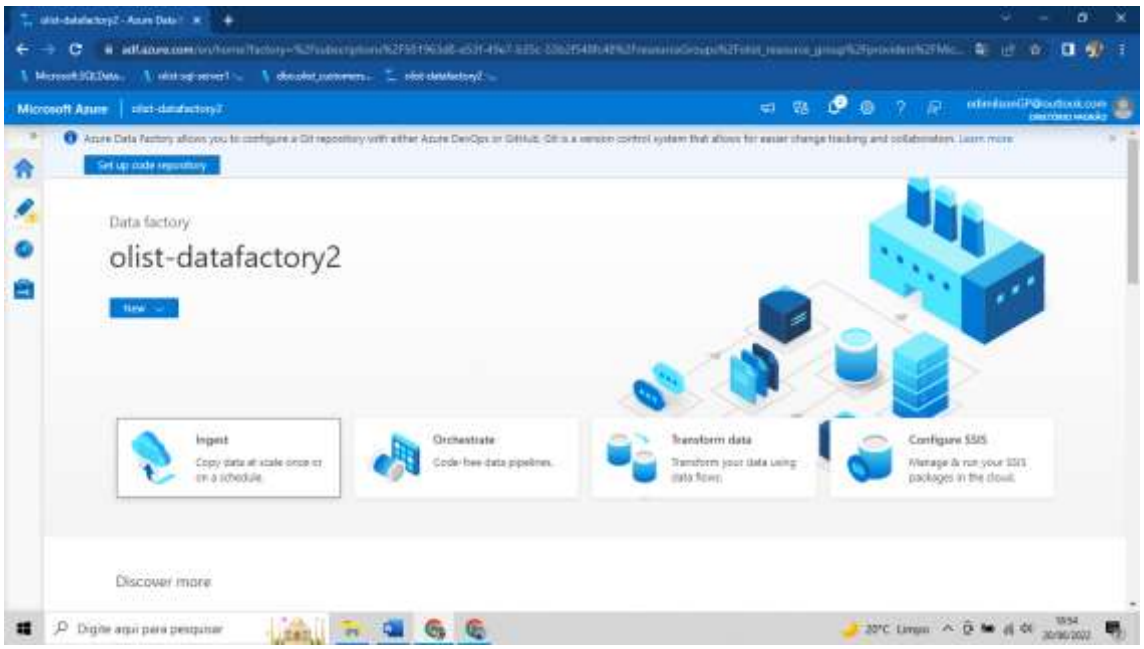
FERRAMENTA PARA CRIAÇÃO MAIS RAPIDA

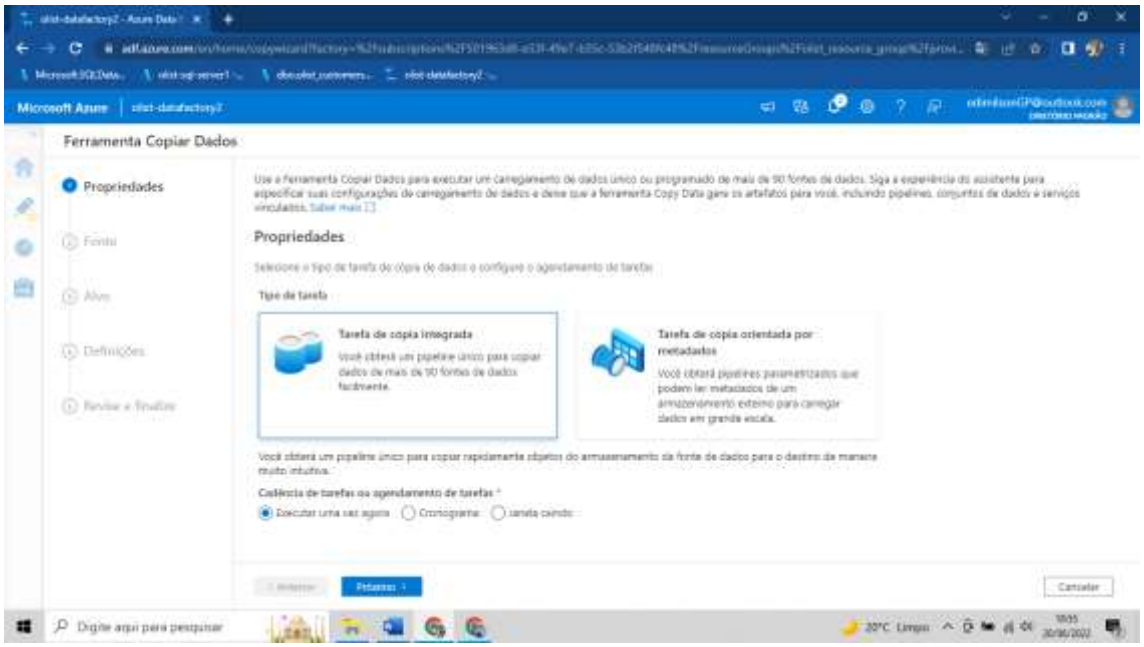
ForEach = ele repete uma tarefa de cada lista que você passa pra ele



Ingestão de dados

CRIANDO PIPELINE PELO WIZARD DO DATA FACTORY





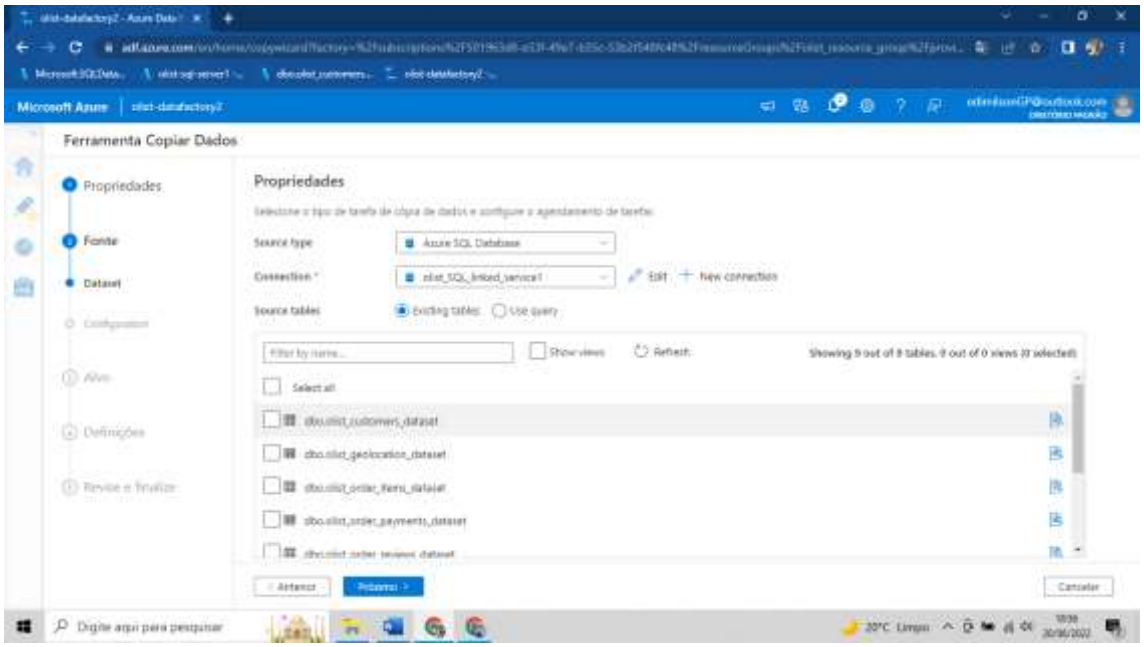
MOSTRANDO TODAS AS TABELAS DISPONIVEIS

SELECIONANDO O BANCO DE DADOS DO SQL COM AZURE

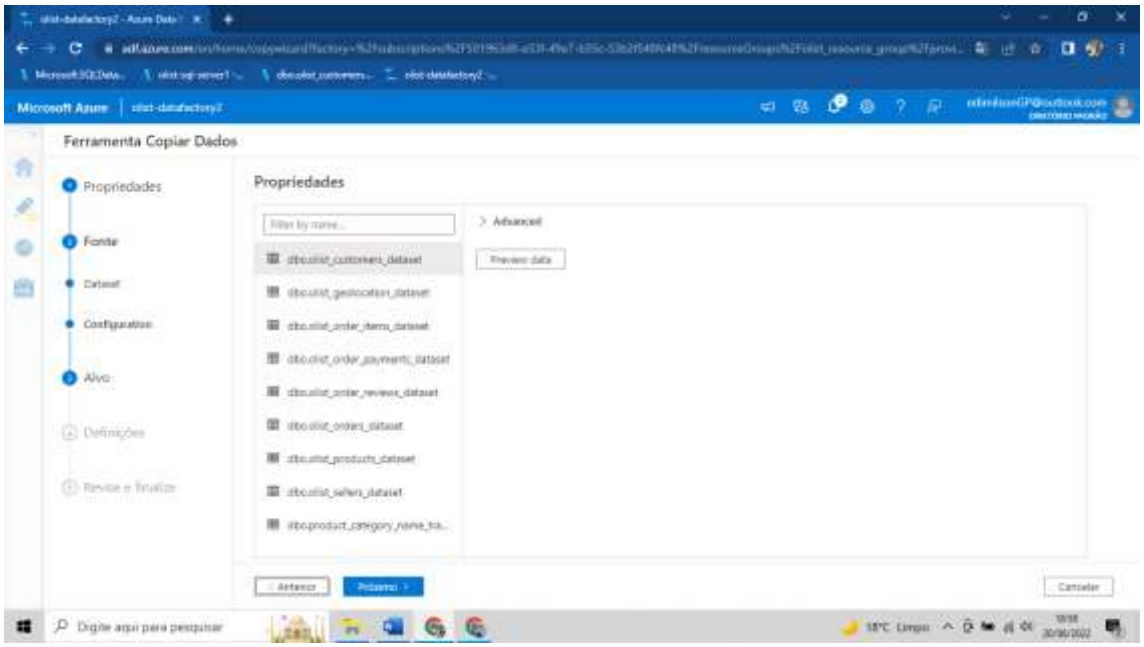
E INFORMANDO A CONEXÃO

ETL

CARGA dos DADOS



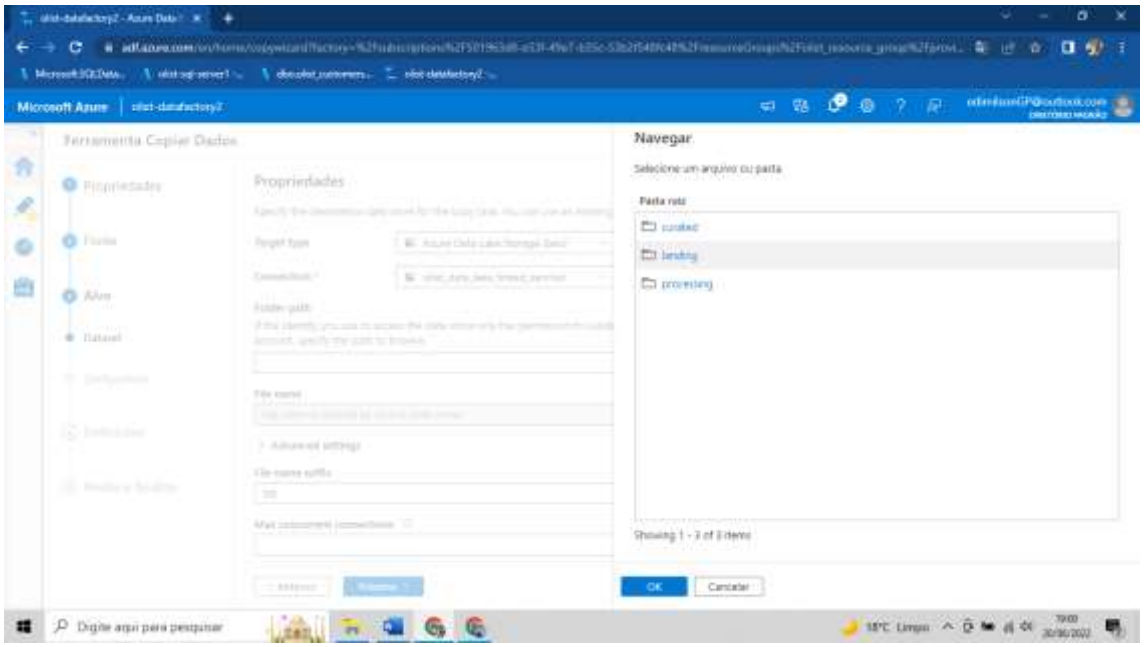
CLICANDO EM PROXIMO



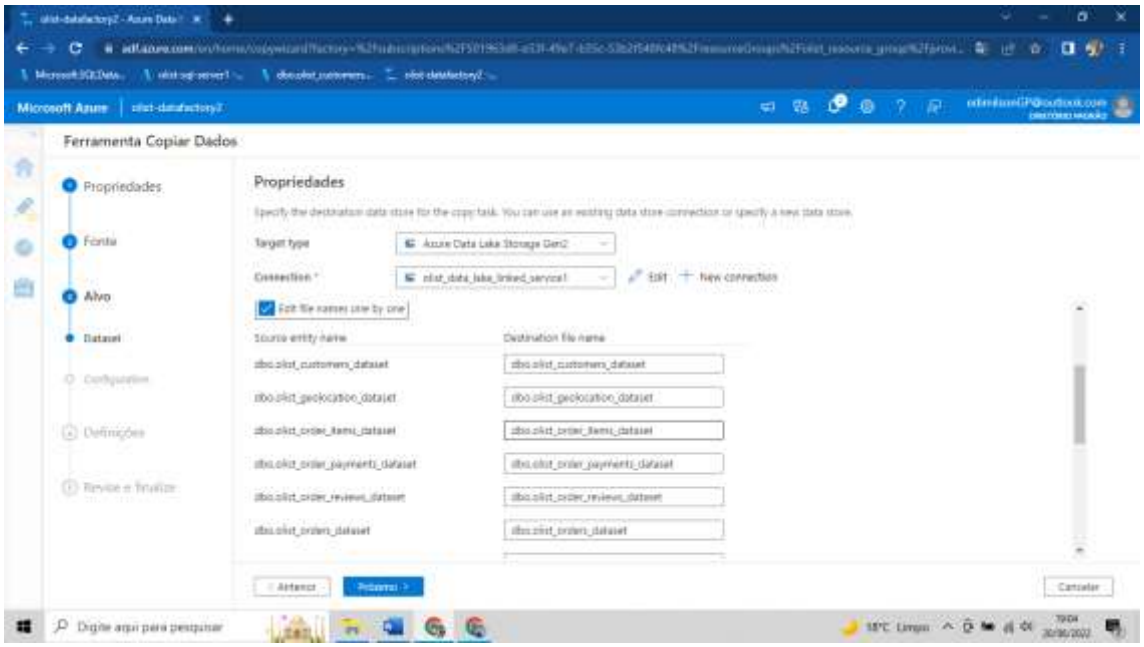
TIPO DE DESTINO

CONEXÃO

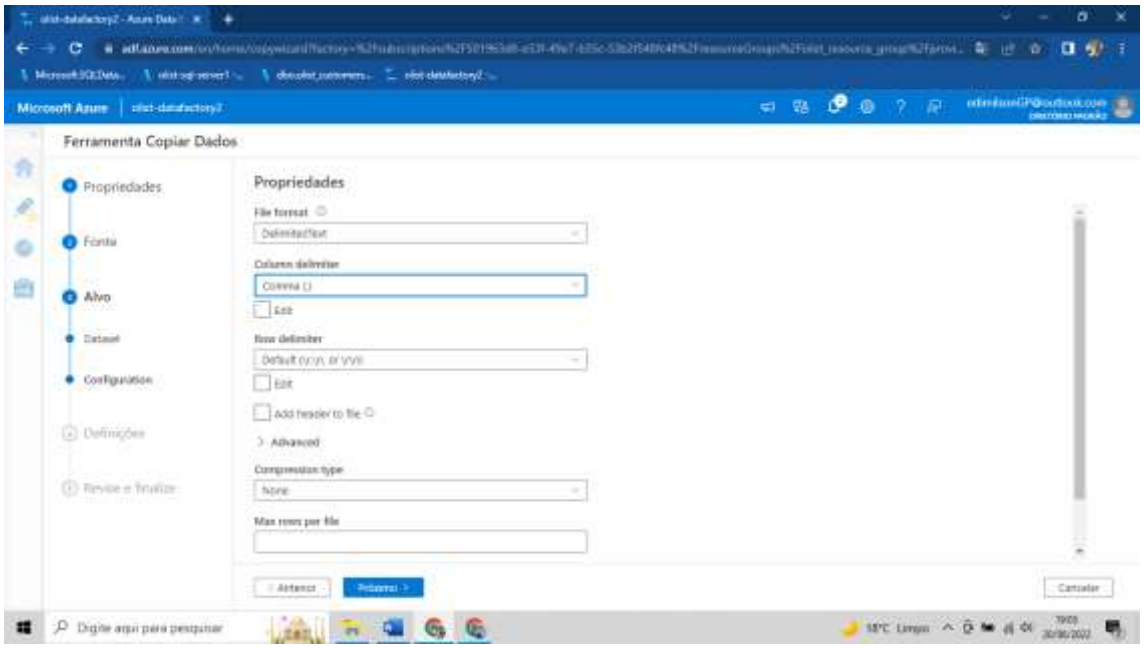
ESCOLHENDO O CONTEINER DE LANDING



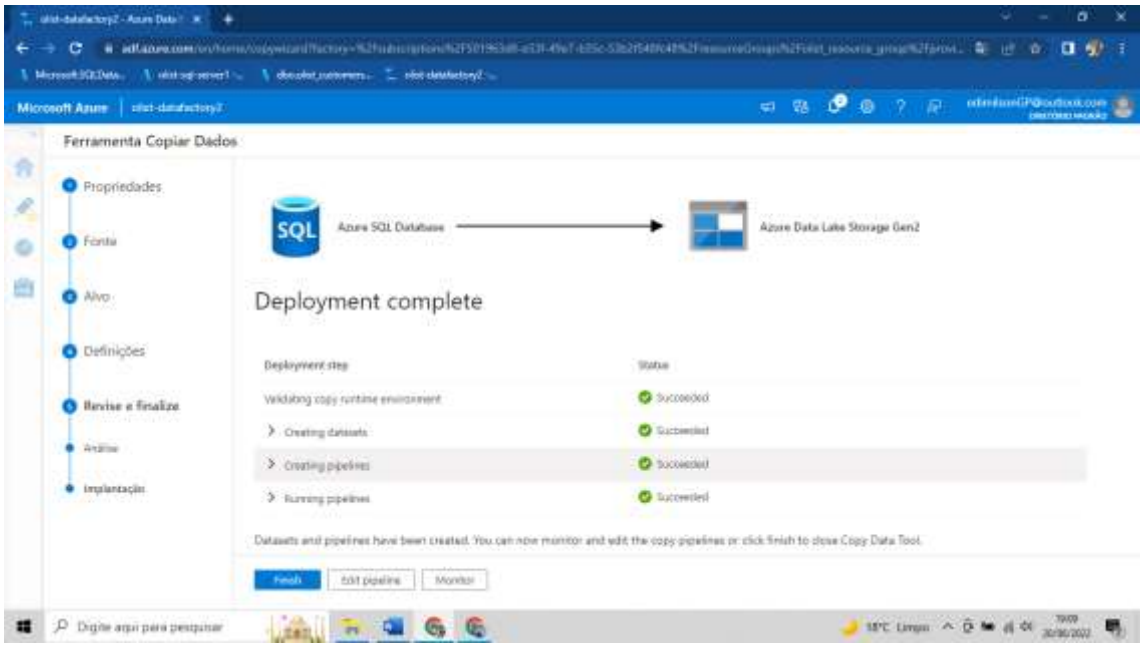
E escolhendo como formato CSV



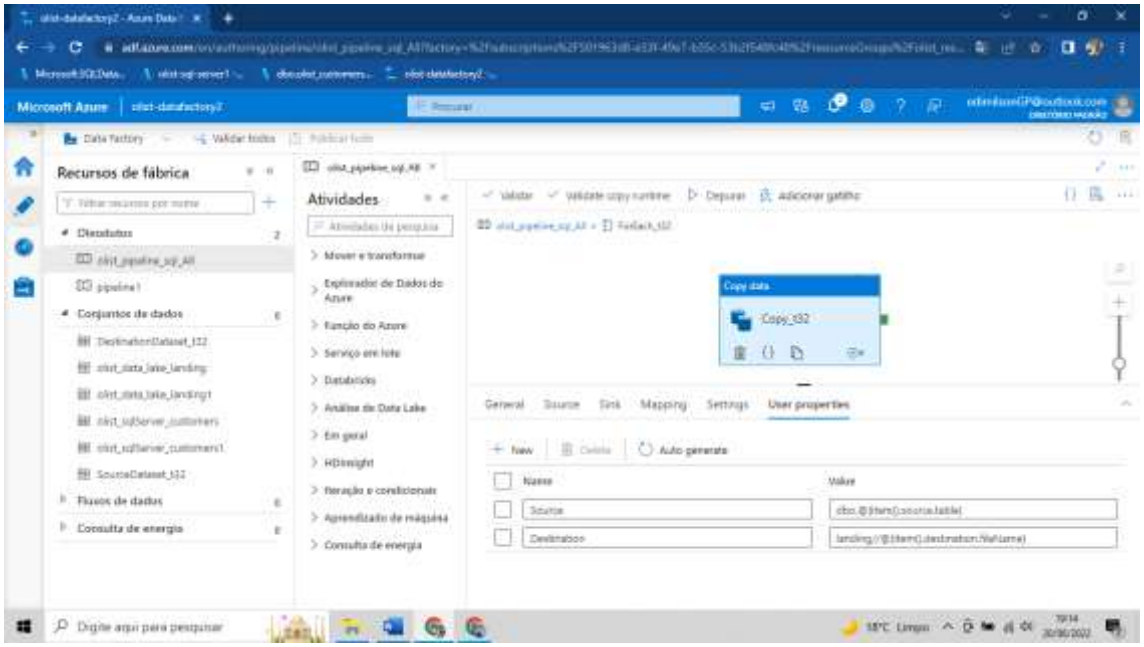
Aqui informa o delimitador que no caso é por virgula se eu quero comprimir o arquivo

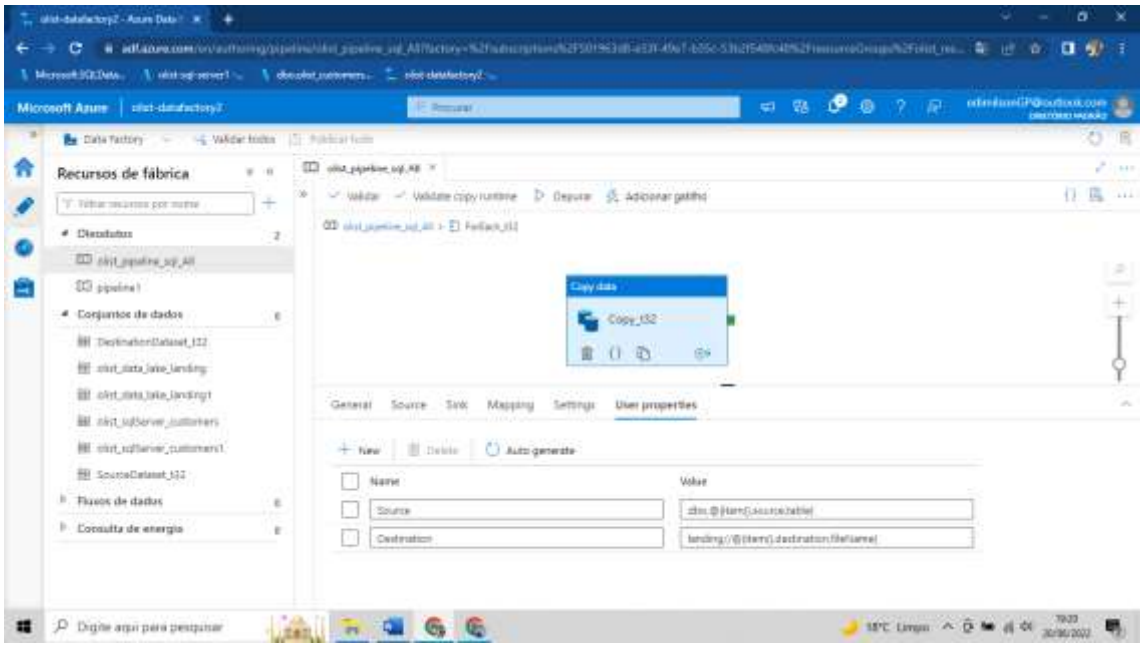


CRIANDO O NOME DO PIPELINE PARA INFORMAR QUE ESTOU ENVIANDO TODAS AS TABELAS



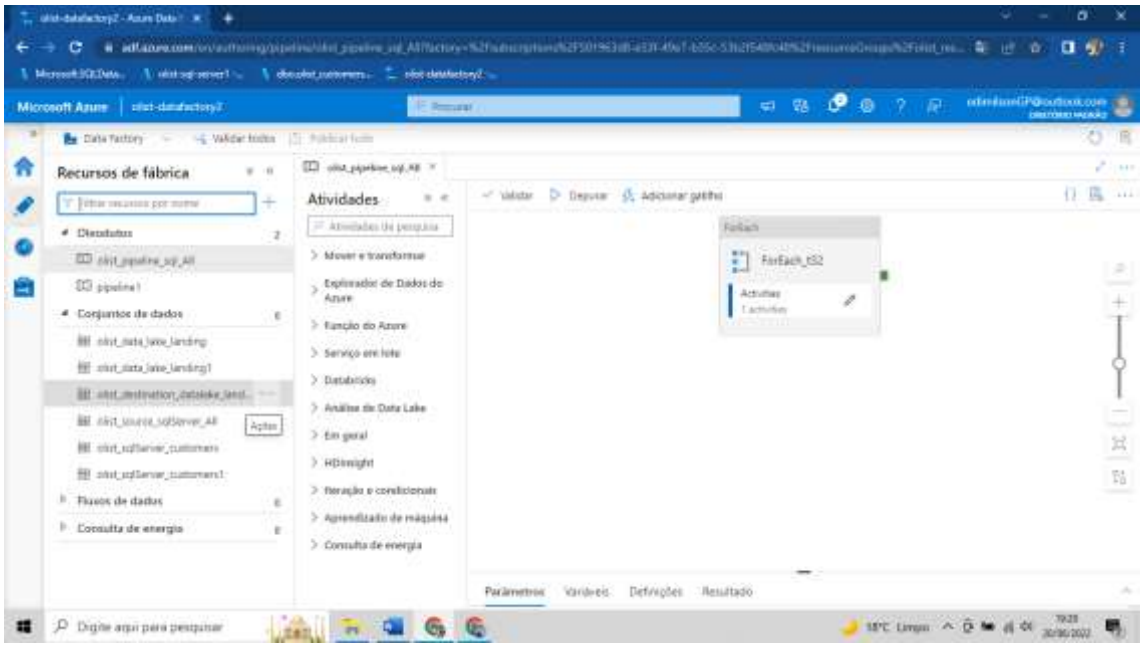
FEZ TODOS O MAPEAMENTO AGORA SÓ EXECUTAR E APERTARCIONAR GATILHO





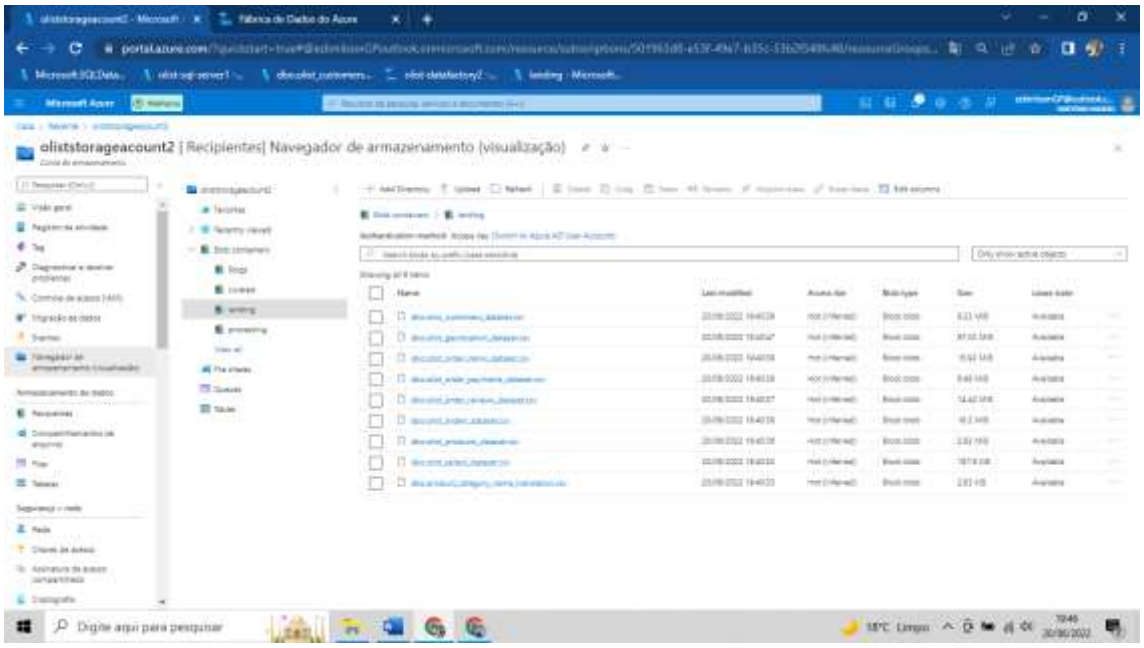
TESTANDO A EXECUÇÃO

Tabelas renomeadas Source e Destination All



Em StorageAcount todas as tabelas foram carregadas em Landing

Tabelas carregadas e gravadas



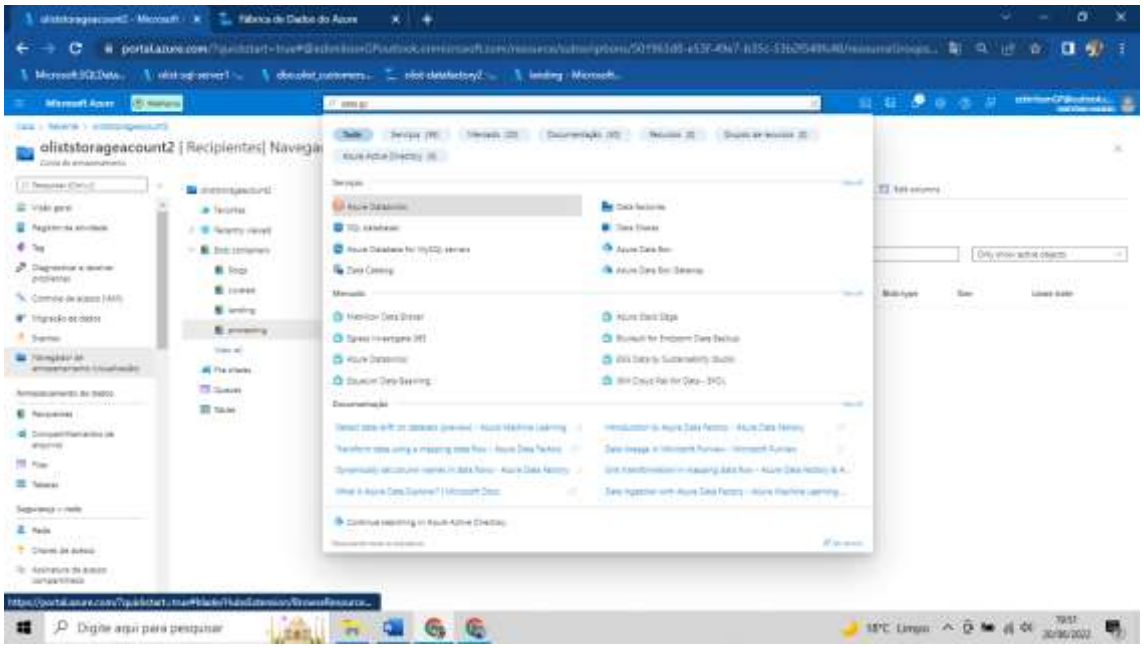
ETL

PROCESSAMENTO = TRANSFORMAÇÃO DOS DADOS

PROCURAR POR AZURE DATABRICKS

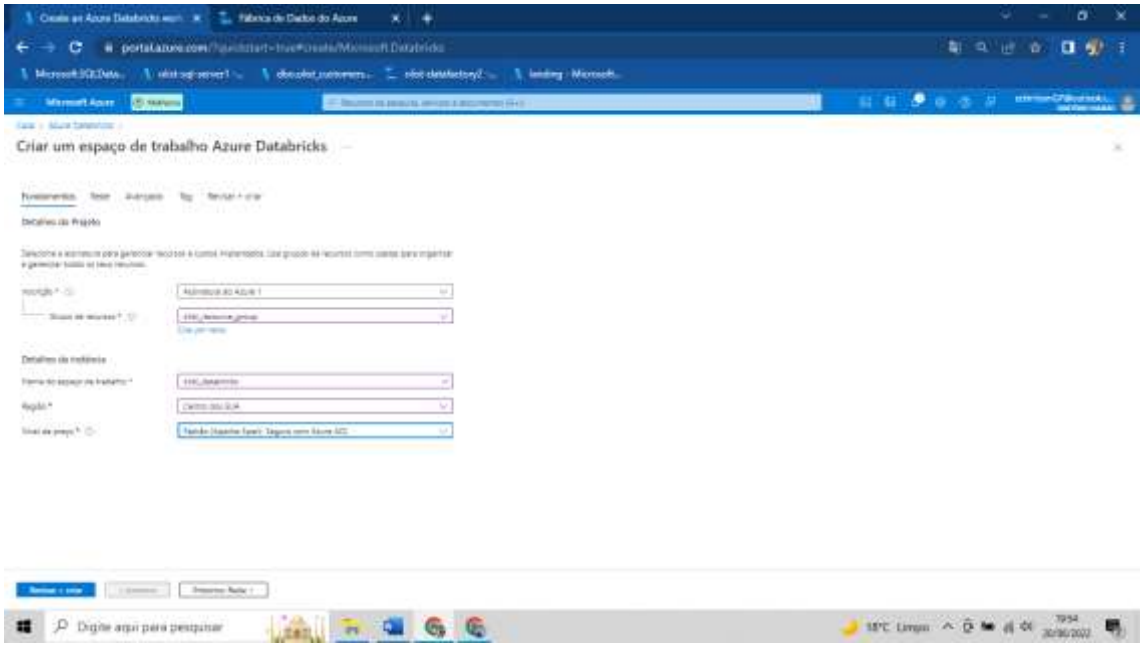
Depois clicar em Criar

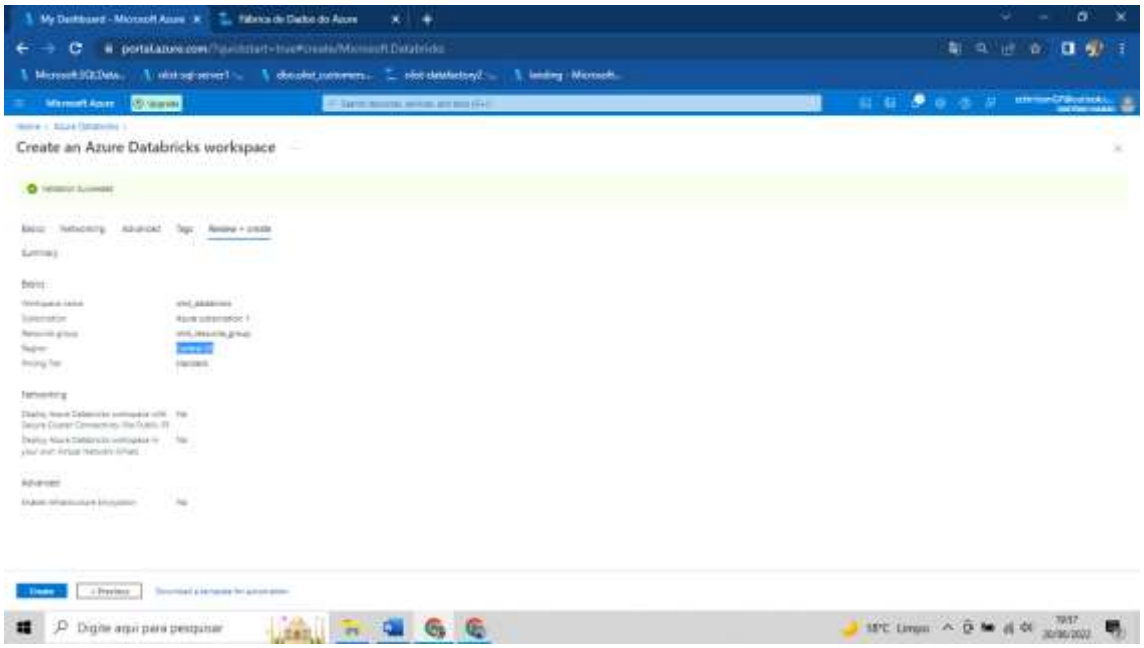
CRIANDO O DATABRICKS



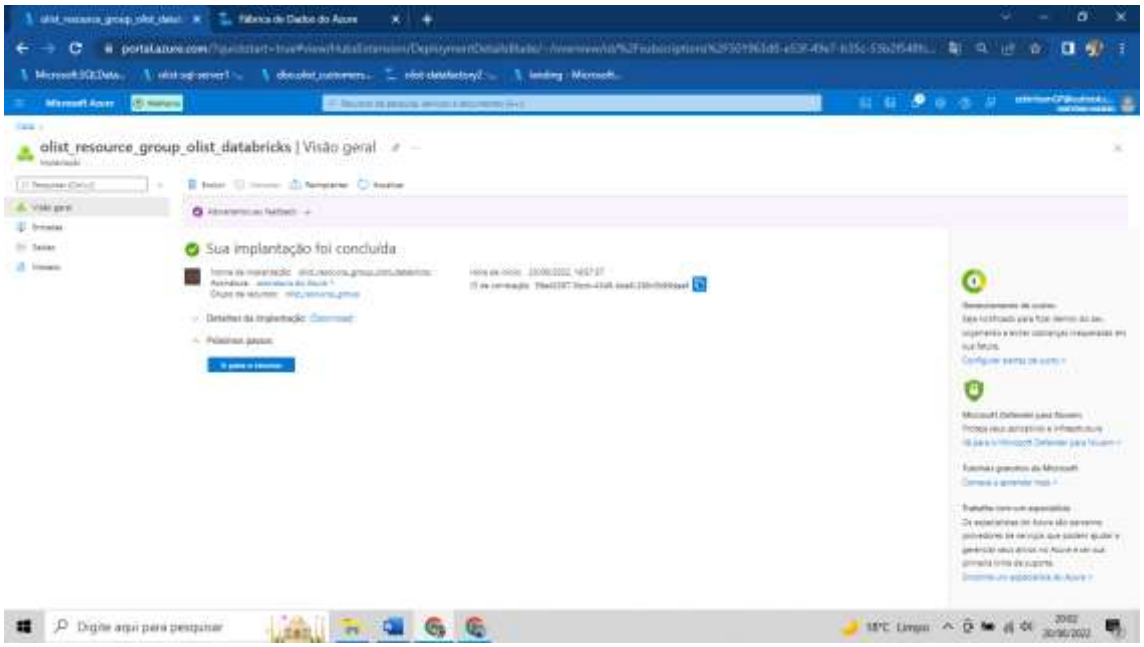
Preenchendo as informações de assinatura e selecionando o grupo de recursos e nome do espaço de trabalho

Região mais barata para o custo com DATABRICKS



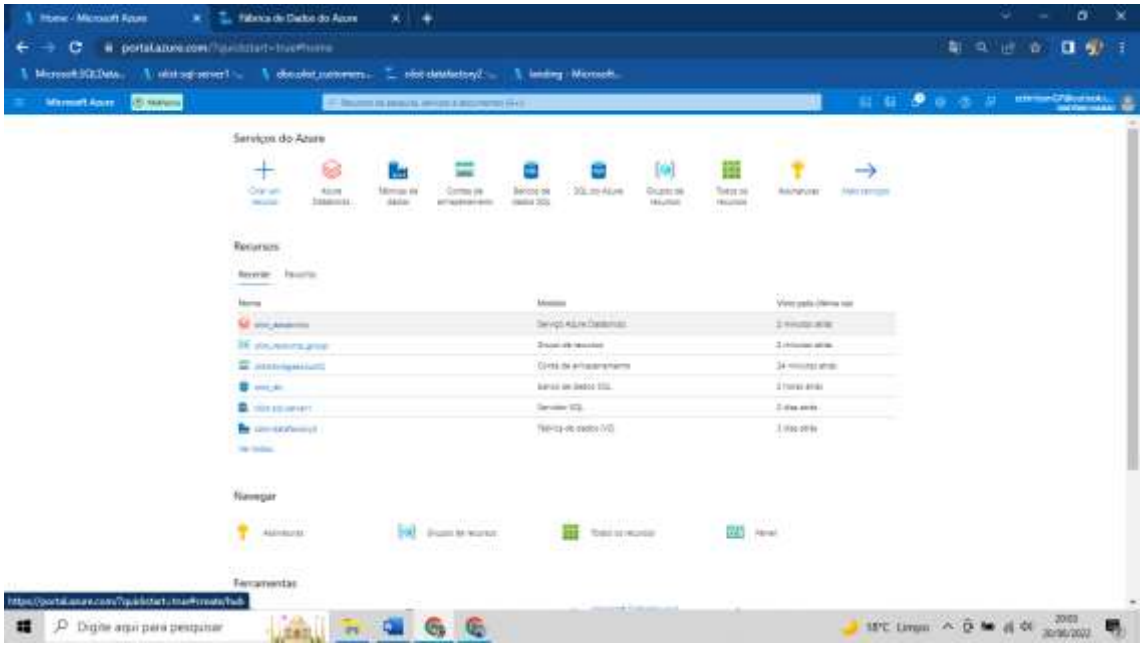


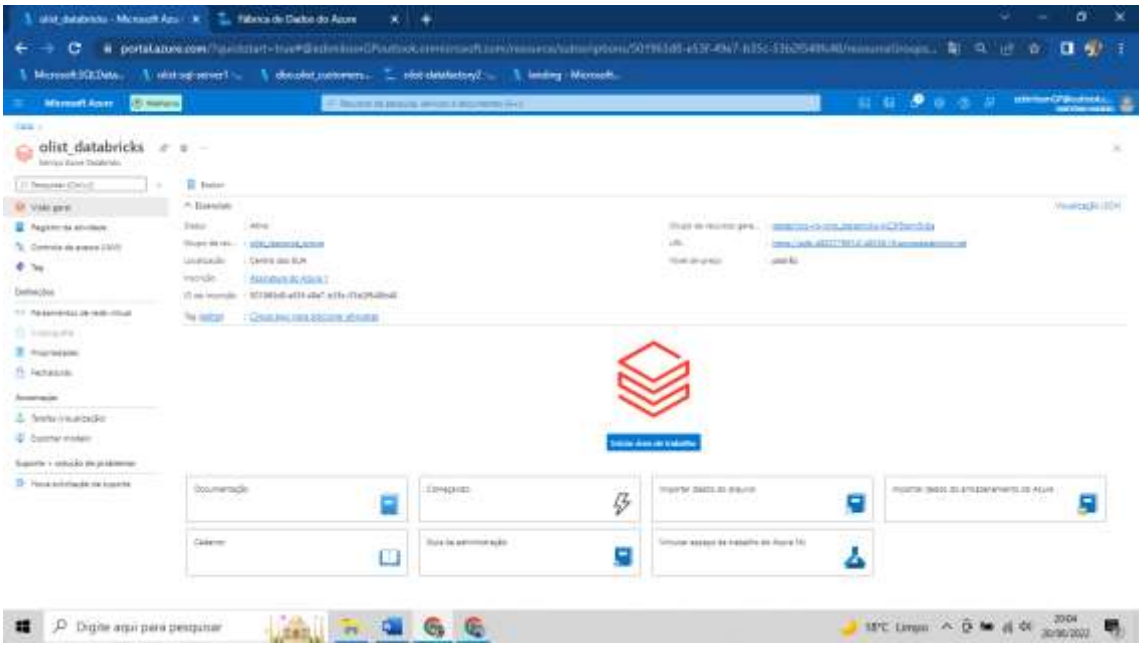
CRIADO O DATABRICKS



VAMOS ACESSAR O RECURSO DO DATABRICKS

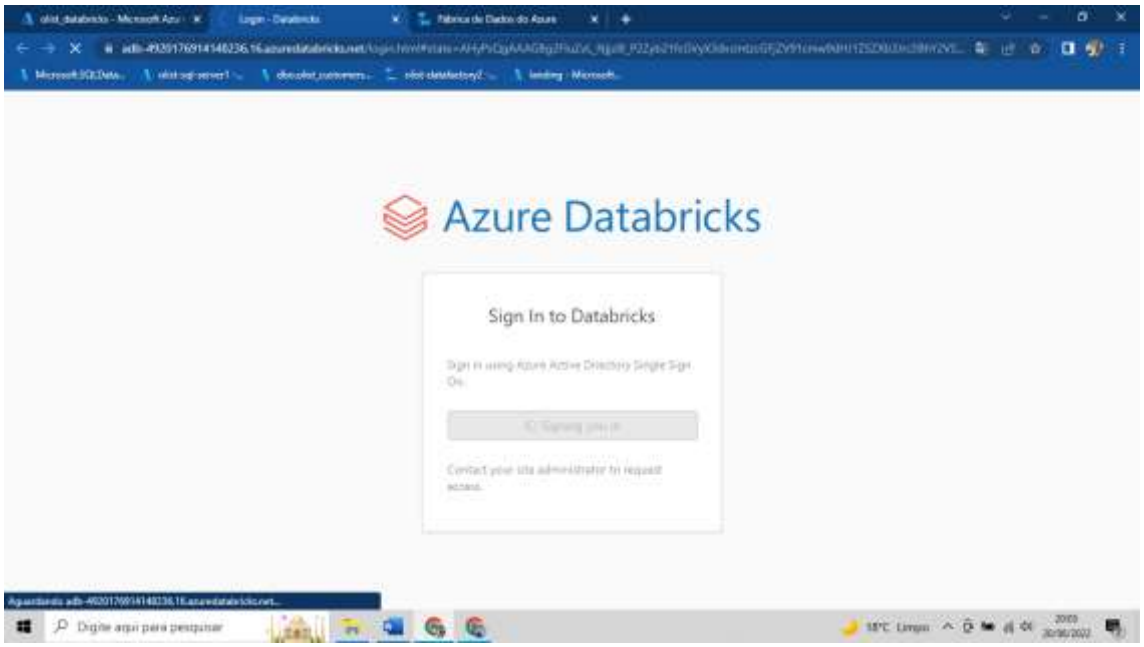
INDO NA PAGINA INICIAL

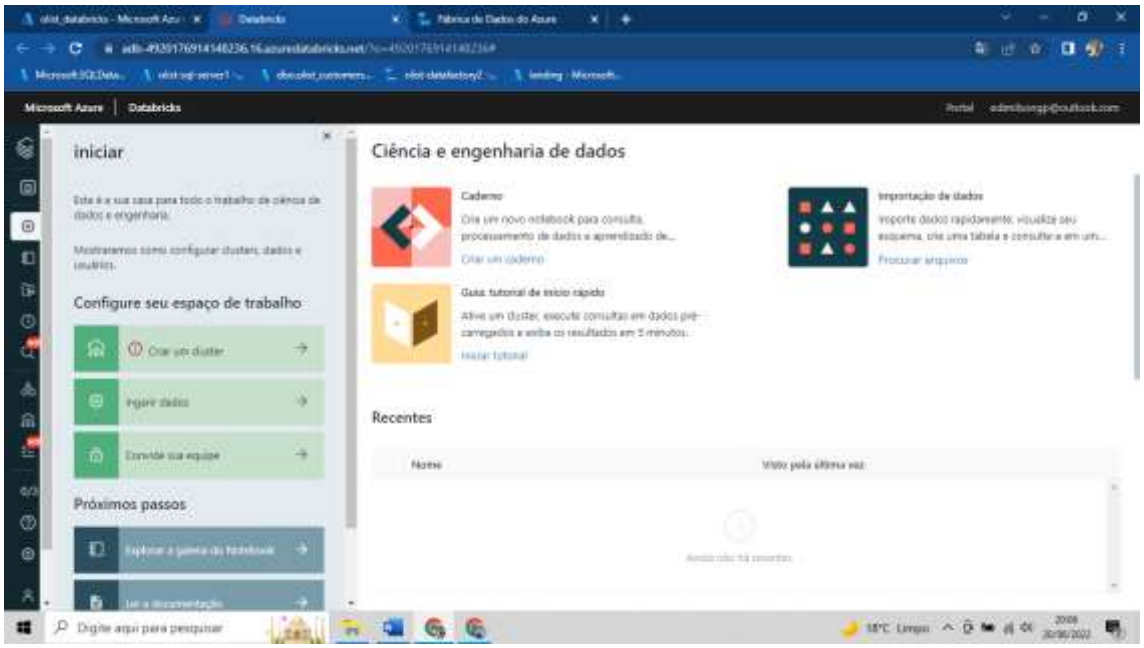




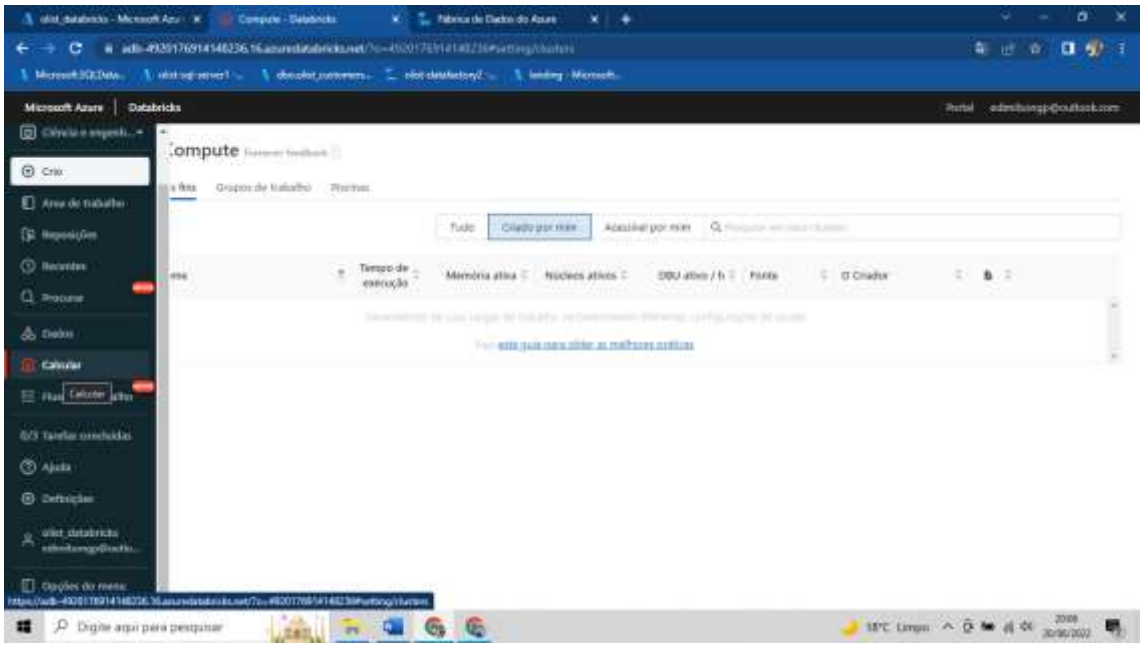
CLICAR EM

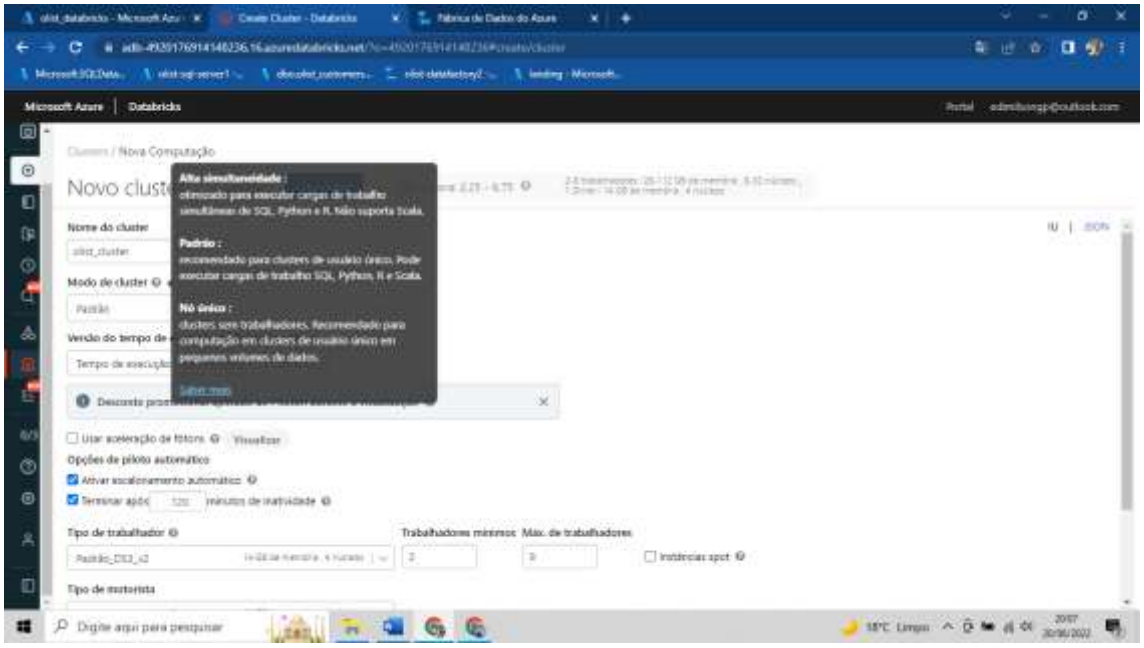
SE CONECTANDO





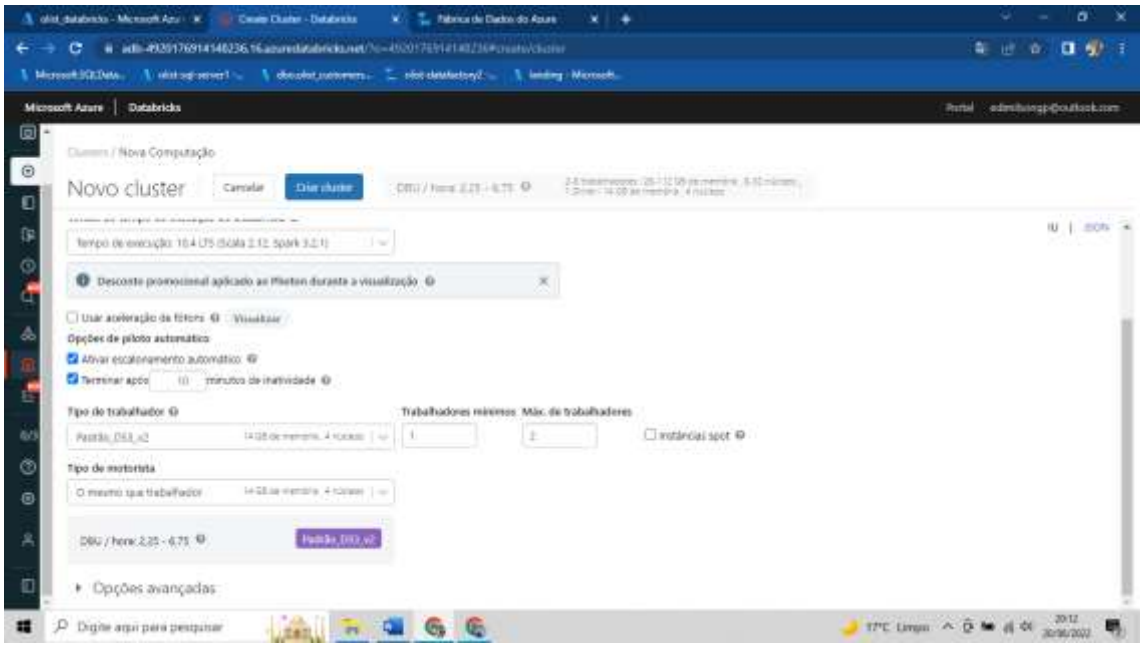
ENTRANDO NO WORKSPACE DO DATABRICKS E CRIANDO UM CLUSTER DE PROCESSAMENTO





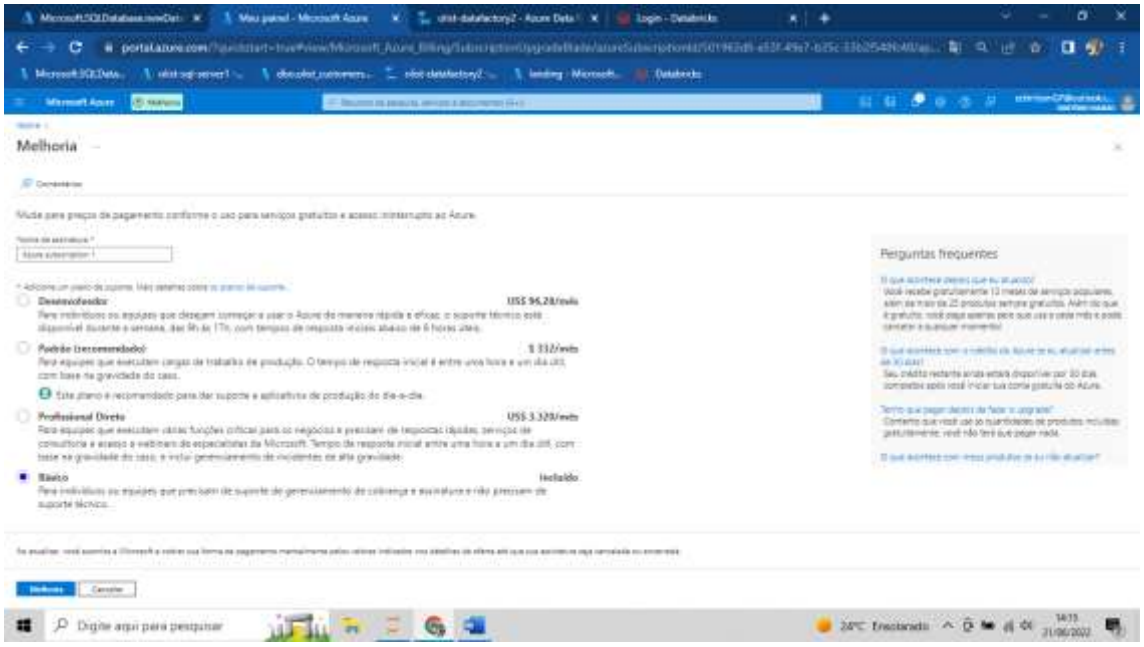
ATENÇÃO PARA O TEMPO

SEMPRE DESATIVAR O CLUSTER

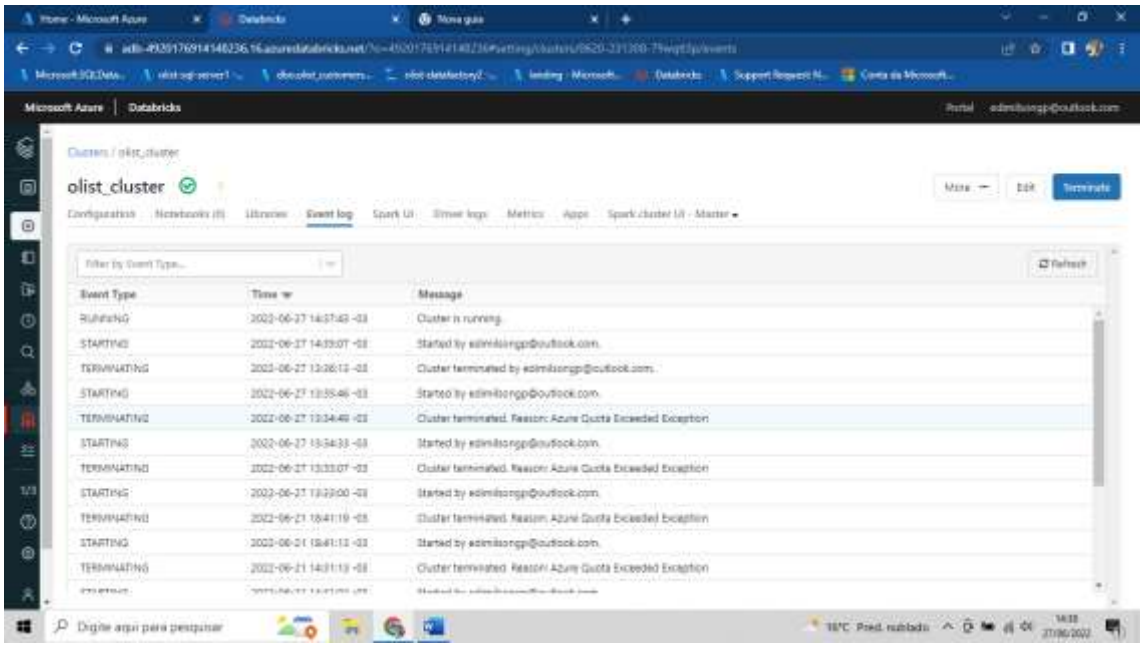


COMO NÃO STARTOU DEVEMOS FAZER UPGRAD DA CONTA PARA PAGO CONFORME O USO, PARA CONSERGUIR UTILIZAR O CLUSTER DO DATABRICKS

FAZER UPGRAD DA CONTA PARA PAGO CONFORME O USO, PARA CONSERGUIR UTILIZAR O CLUSTER DO DATABRICKS



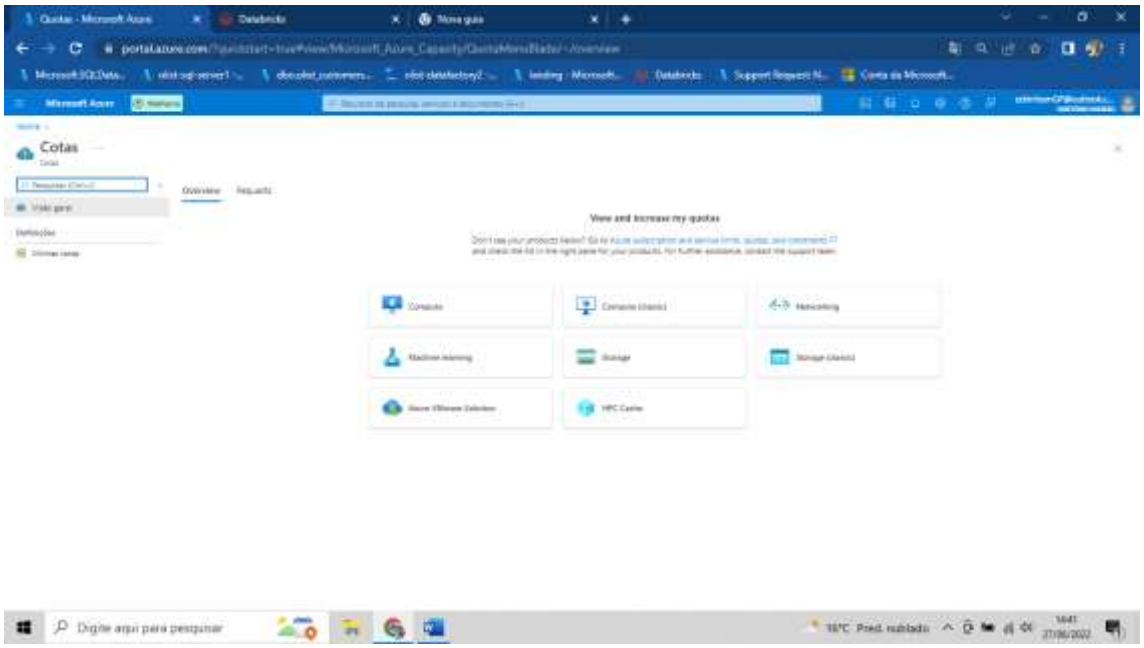
CLUSTER CRIADO E ASSINATURA ATIVADA



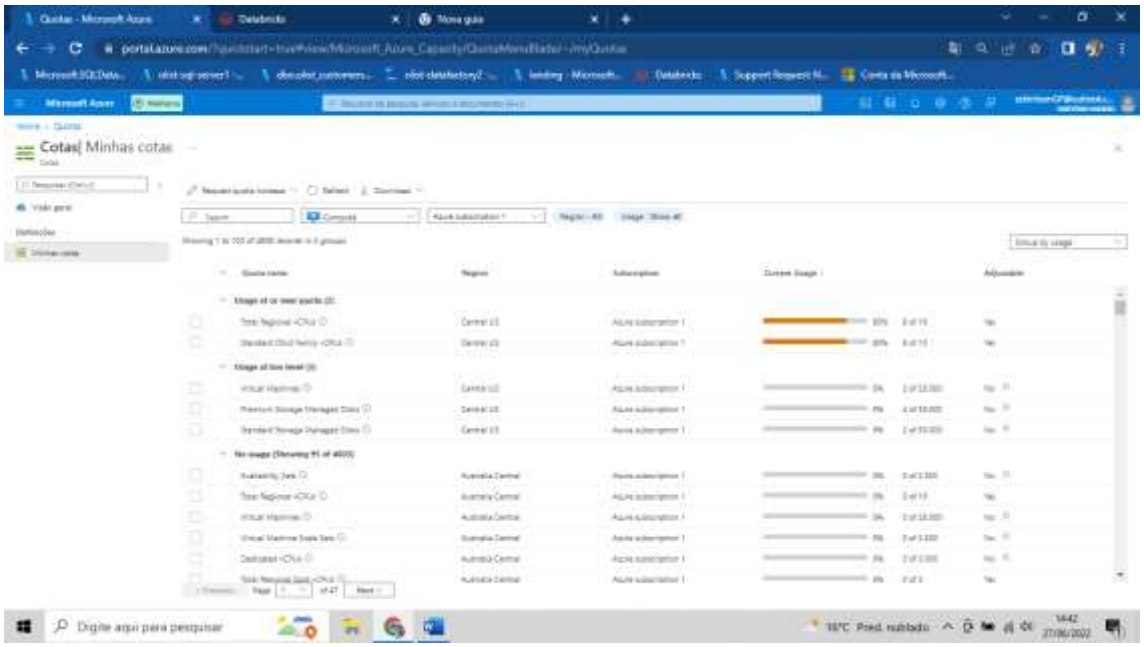
CASO MEU CLUSTER NÃO SUBA CORRETAMENTE IREMOS SOLICITAR O AUMENTO DE COTAS

CASO TENHA SUBIDO DE FORMA CORRETA NÃO PRECIS, MAS VAMOS APRENDER COMO FAZER.

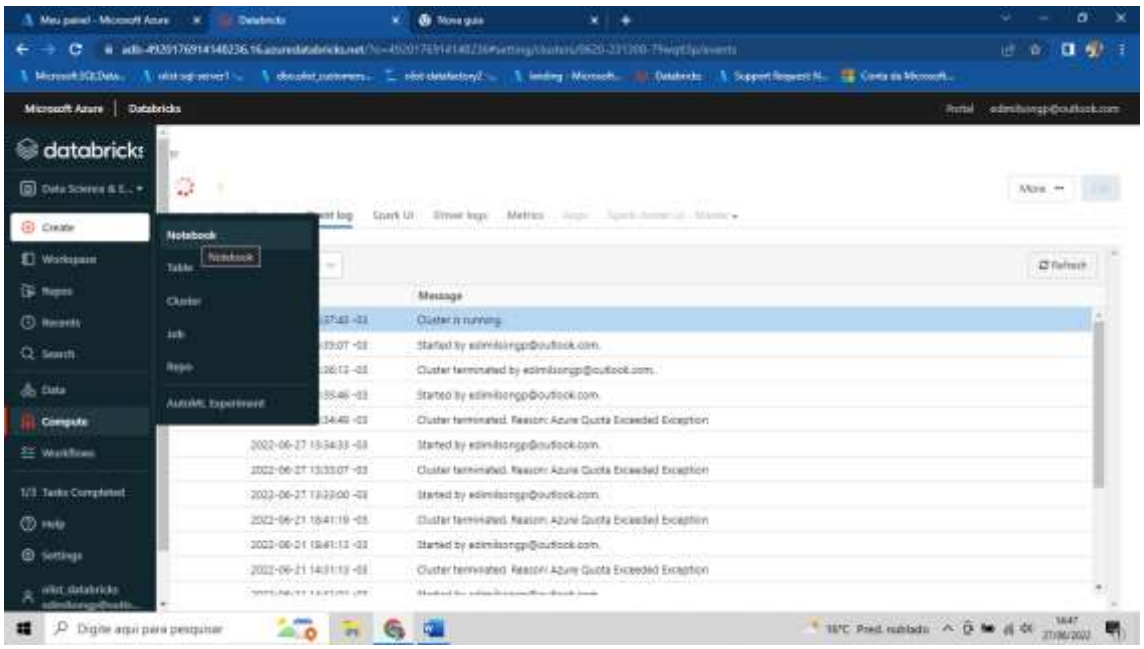
PESQUISAR POR COTA



AUMENTO DE CORE ENTRAR EM COMPUTE



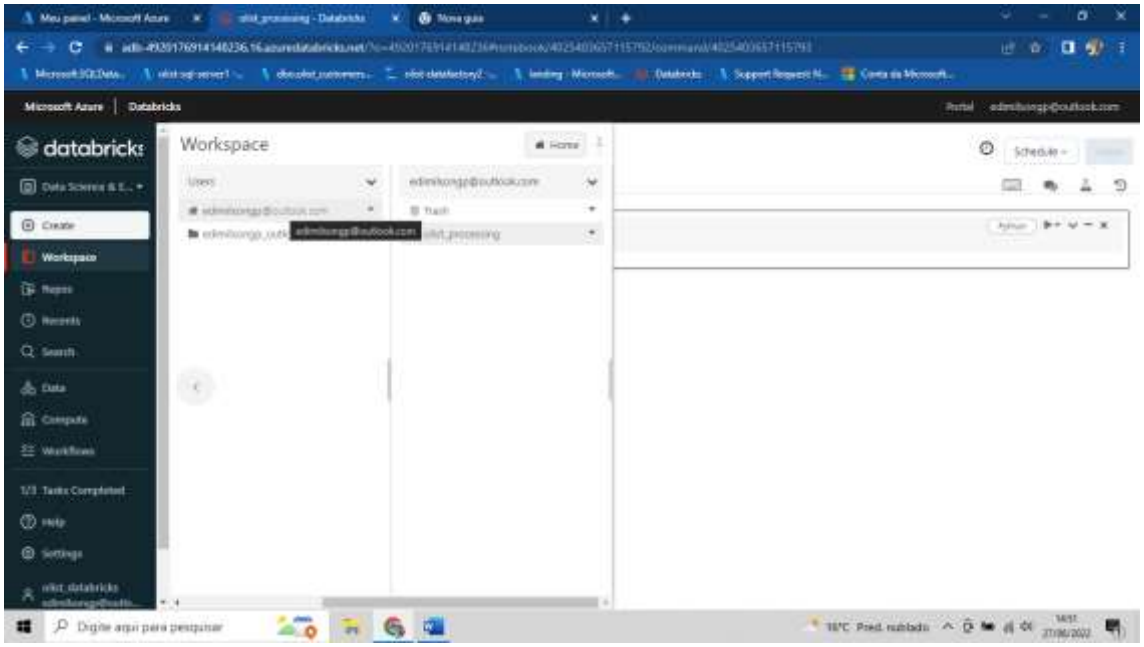
CRIANDO NOTEBOOK PARA PROCESSAMENTO EM PYSPARK

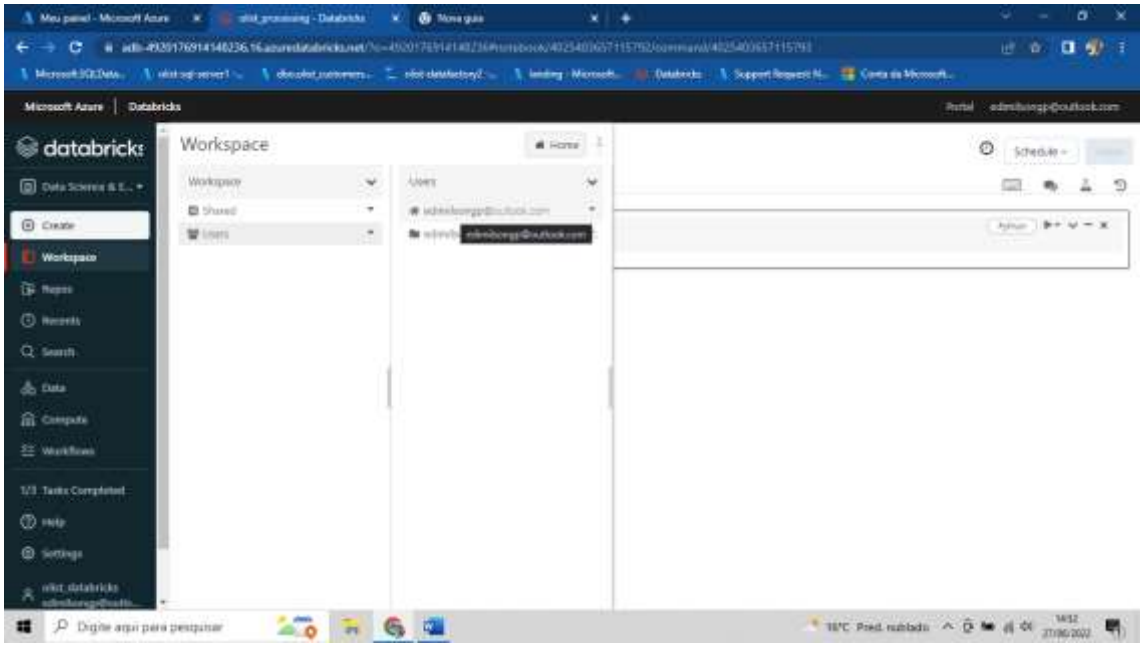


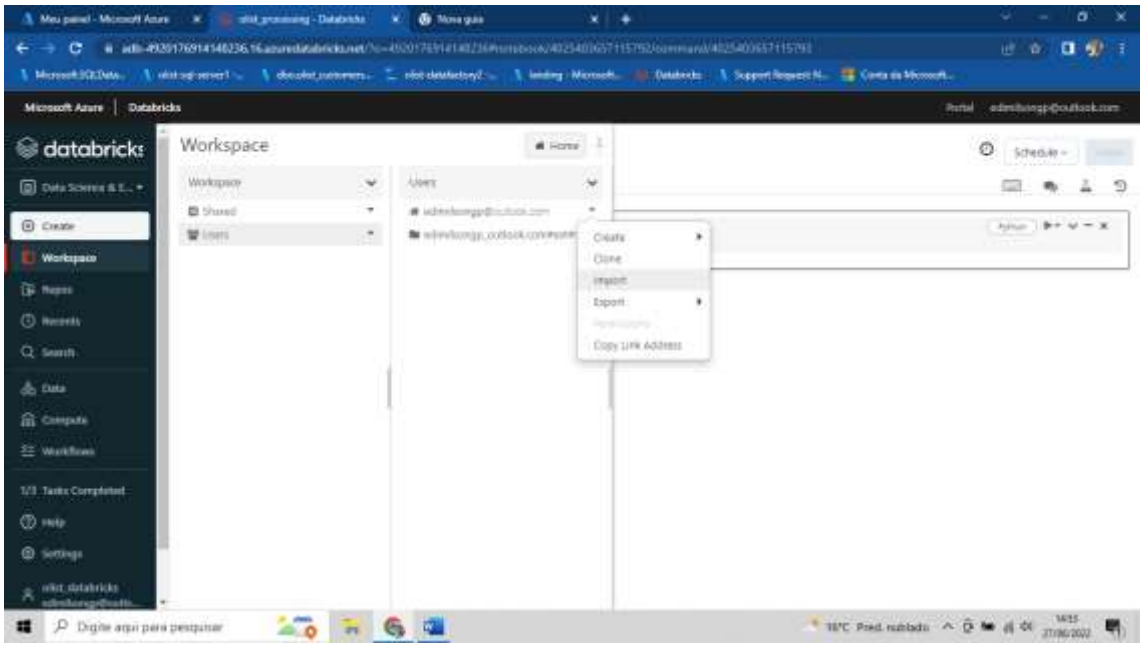
COLOCANDO TODAS AS INFORMAÇÕES E CRIANDO

NOTEBOOK ATIVO EM PYSPARK

IMPORTANDO O SCRIPT







SCRIP IMPORTADO

PASSO A PASSO DO QUE O SCRIPT FAZ

CONFIGURAÇÃO A SEGUIR CONSTA NA DOCUMENTAÇÃO, NECESSARIO APENAS ADAPTAR AS INFORMAÇÕES AOS SEU ARQUIVOS.

#Carregar, Transformar, Persistir Pipeline

#1 - monte os data lakes

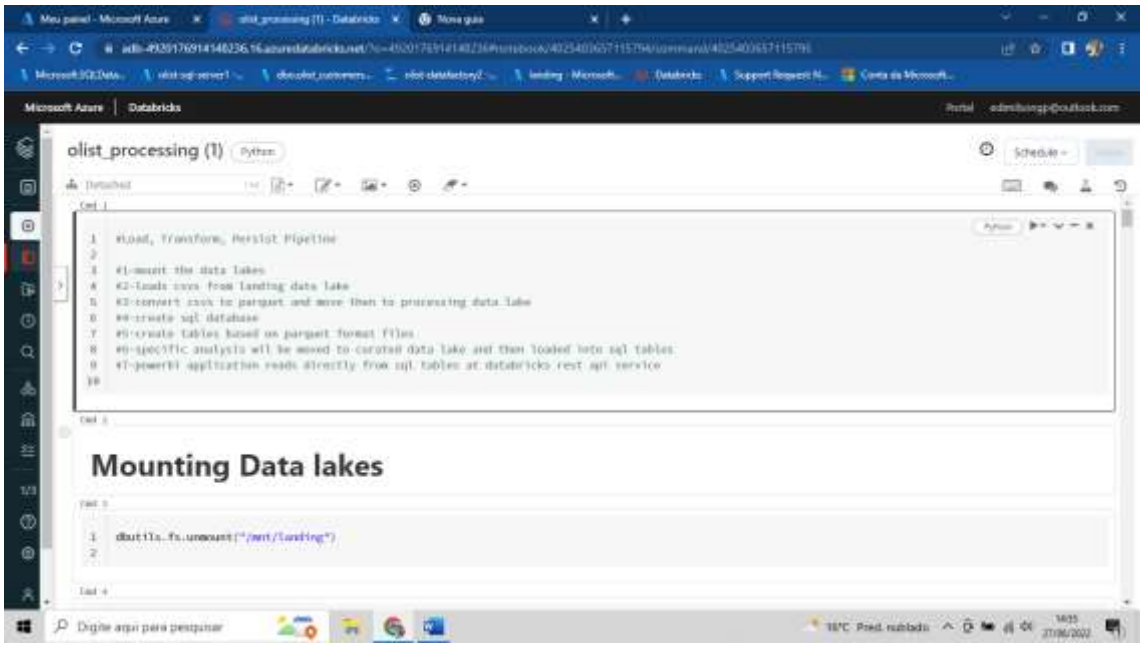
#2-carrega csvs do data lake de desembarque

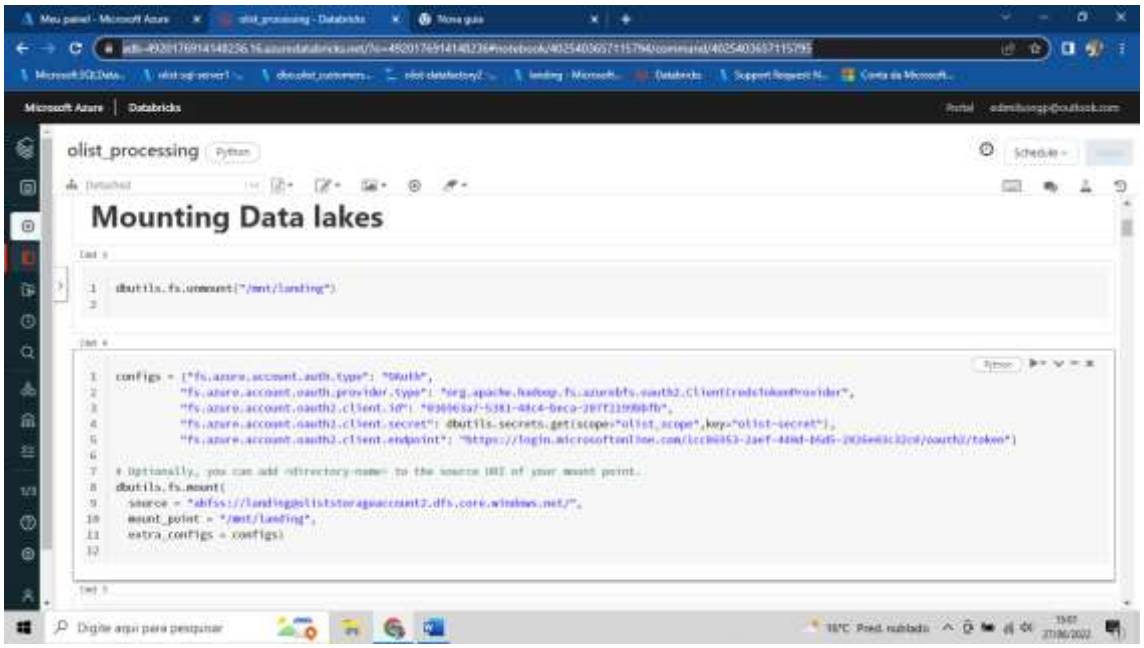
#3-converter csvs para parquet e passar para o data lake de processamento

#4-criar banco de dados sql

#5-crie tabelas com base em arquivos de formato parquet

A análise específica nº 6 será movida para o data lake com curadoria e, em seguida, carregada em tabelas sql

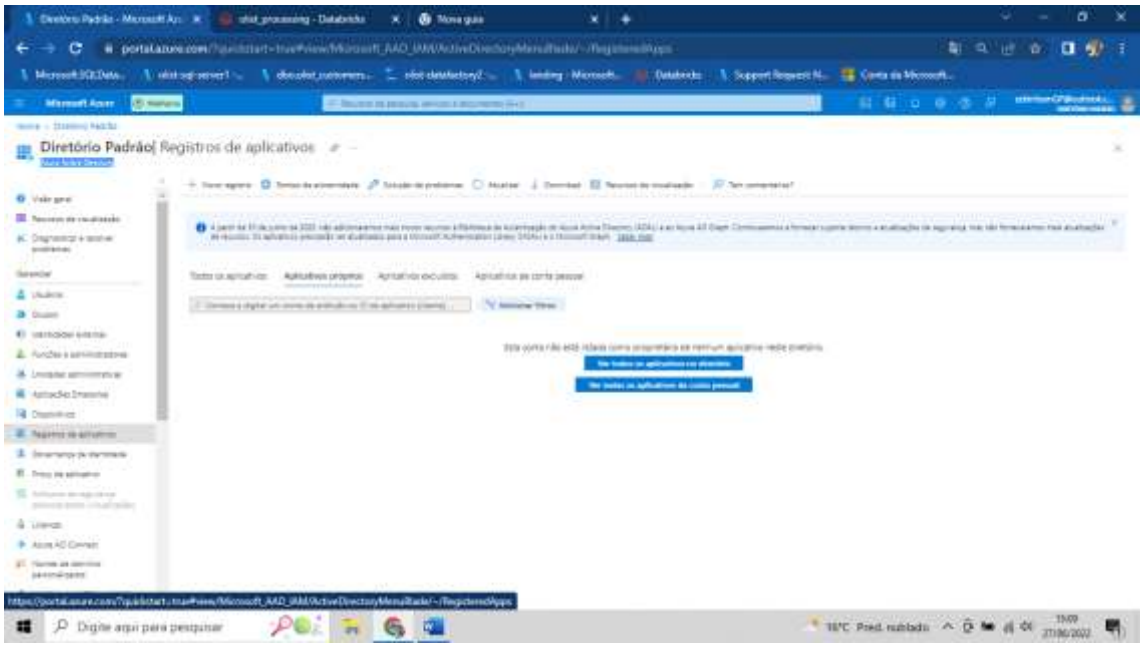


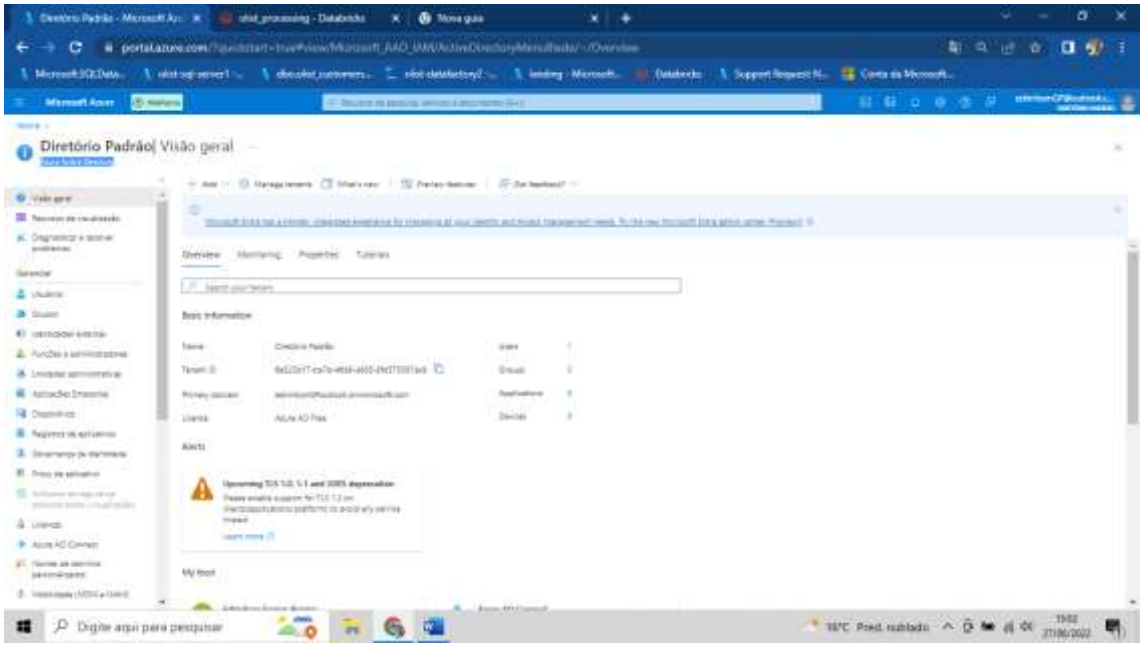


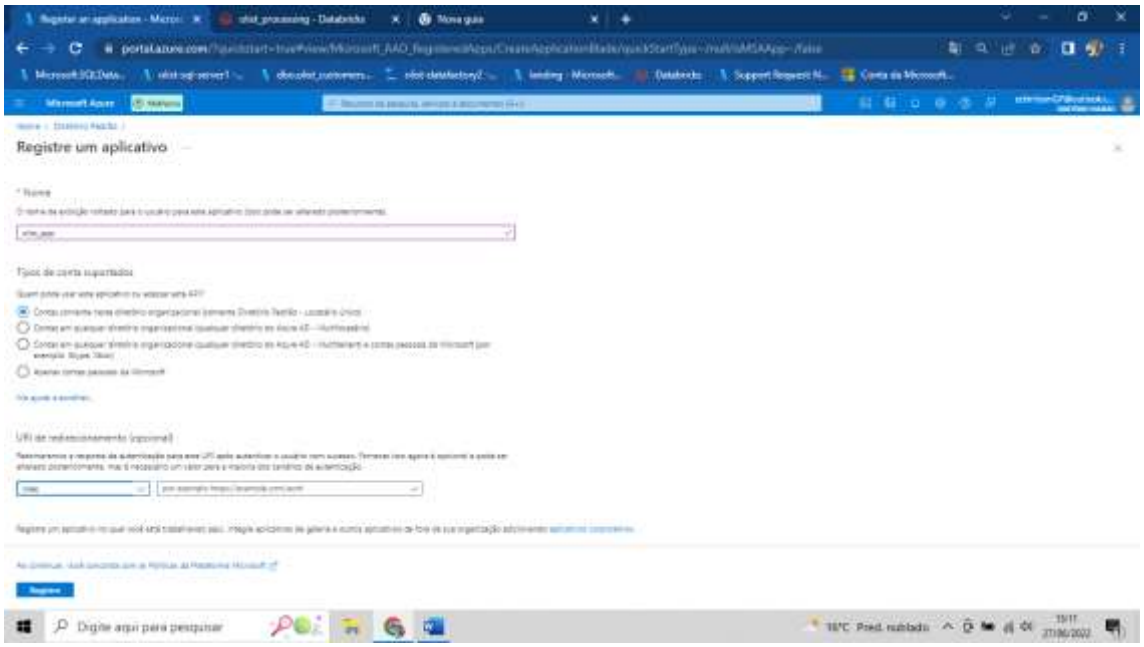
MONTAR O DATA LAKE COM PYSPARK NO DATABRICKS

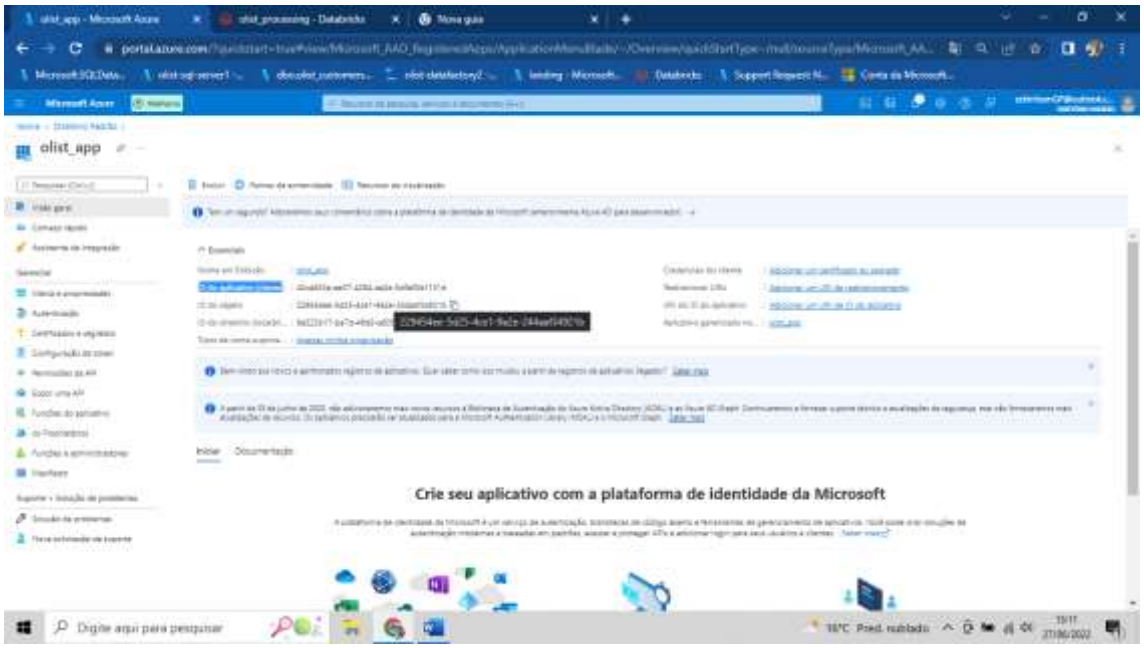
PRECISAMOS REGISTRAR UMA APLICAÇÃO

Azure Active Directory


Diretório Padrão| Registros de aplicativos
Digite aqui para pesquisar

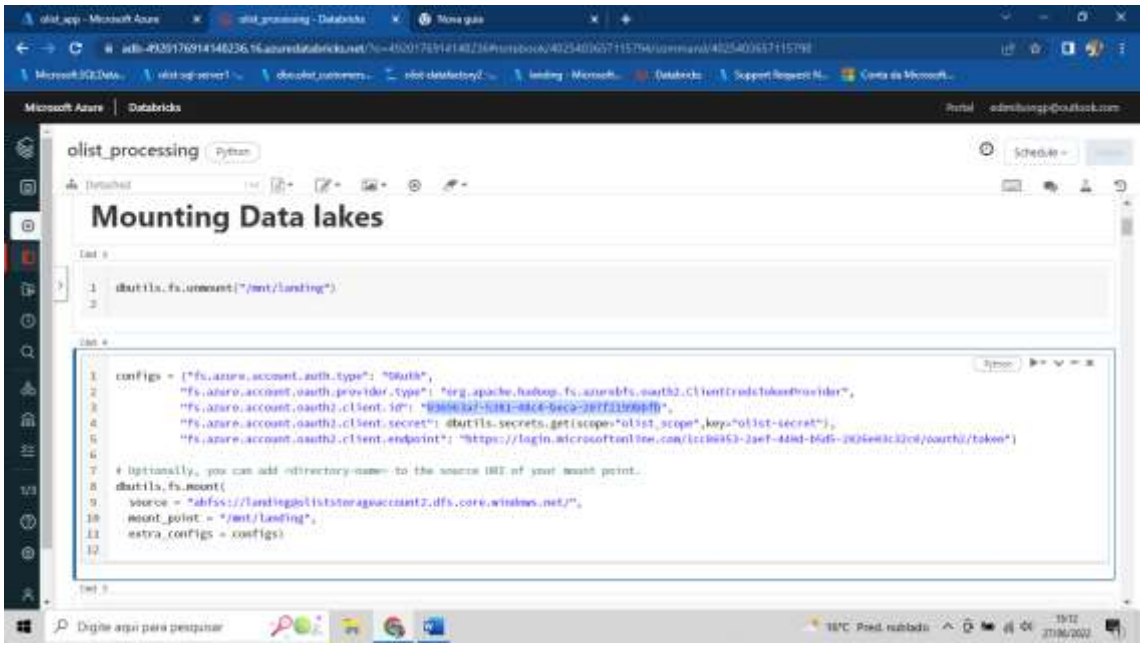

Diretório Padrão| Visão geral
Digite aqui para pesquisar

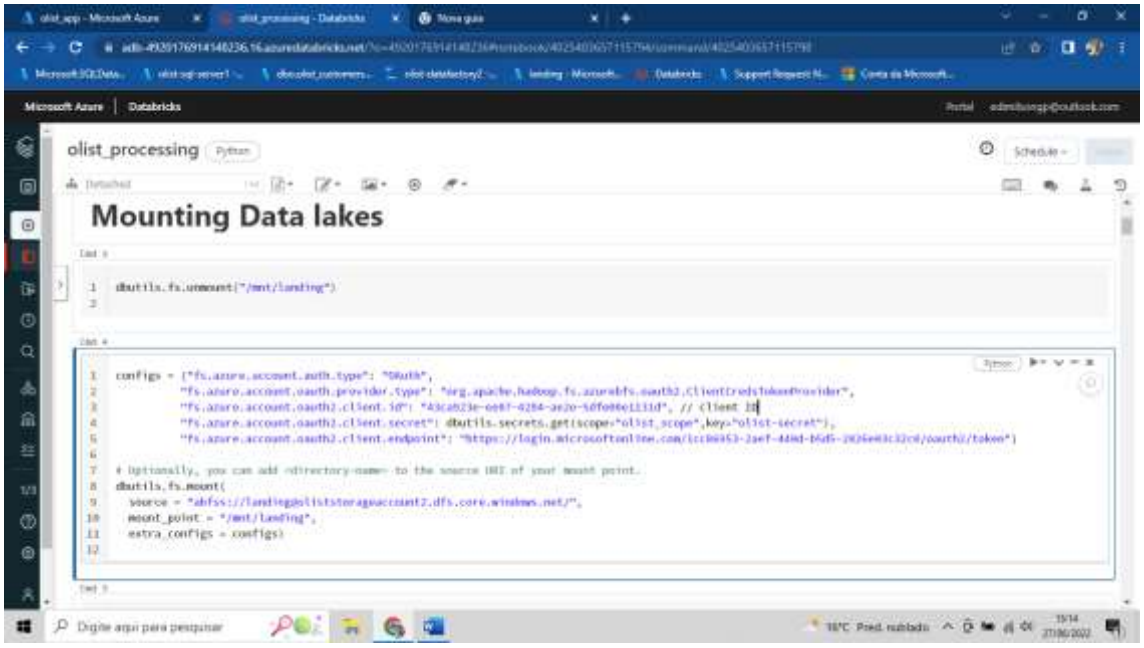




Copiar 43ca923e-ee07-4284-ae2e-5dfe00e1131d

E colar no lugar do 036963a7-5381-48c4-beca-207f2199bbfb

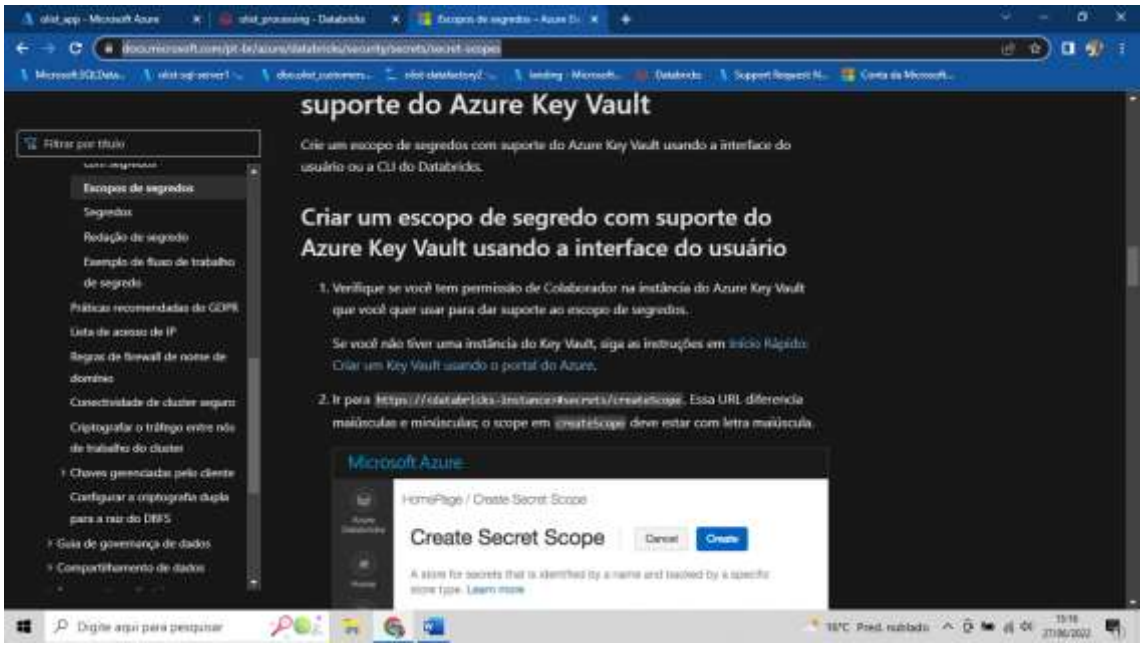




https://docs.microsoft.com/pt-br/azure/databricks/security/secrets/secret-scopes

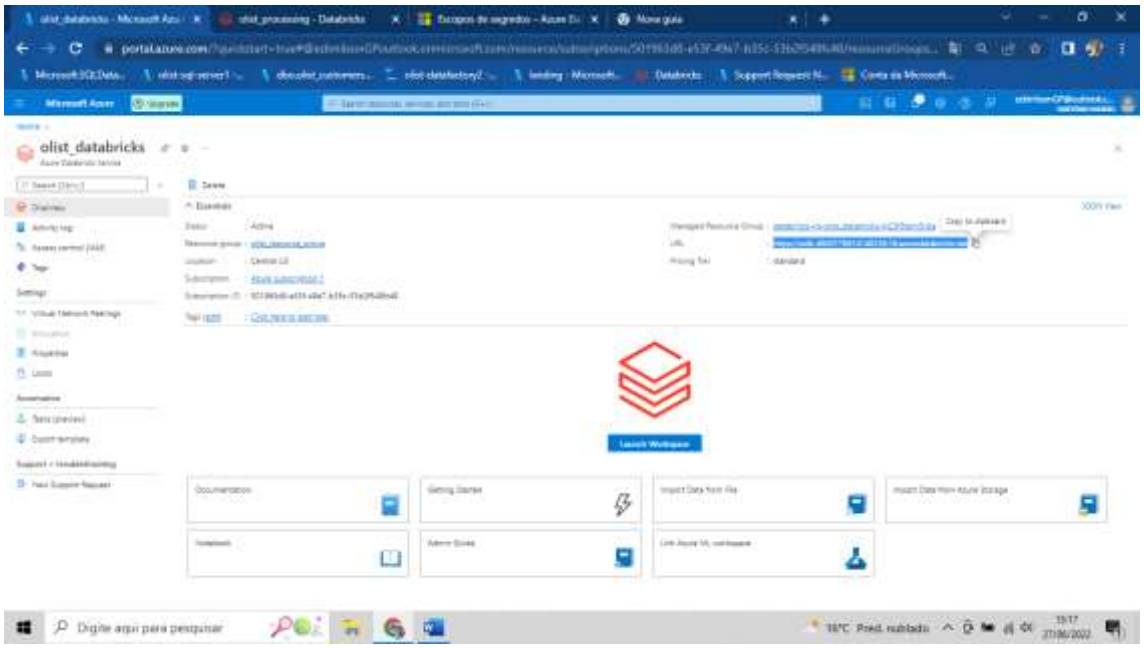
URL DA DOCUMENTAÇÃO

https://<databricks-instance>#secrets/createScope



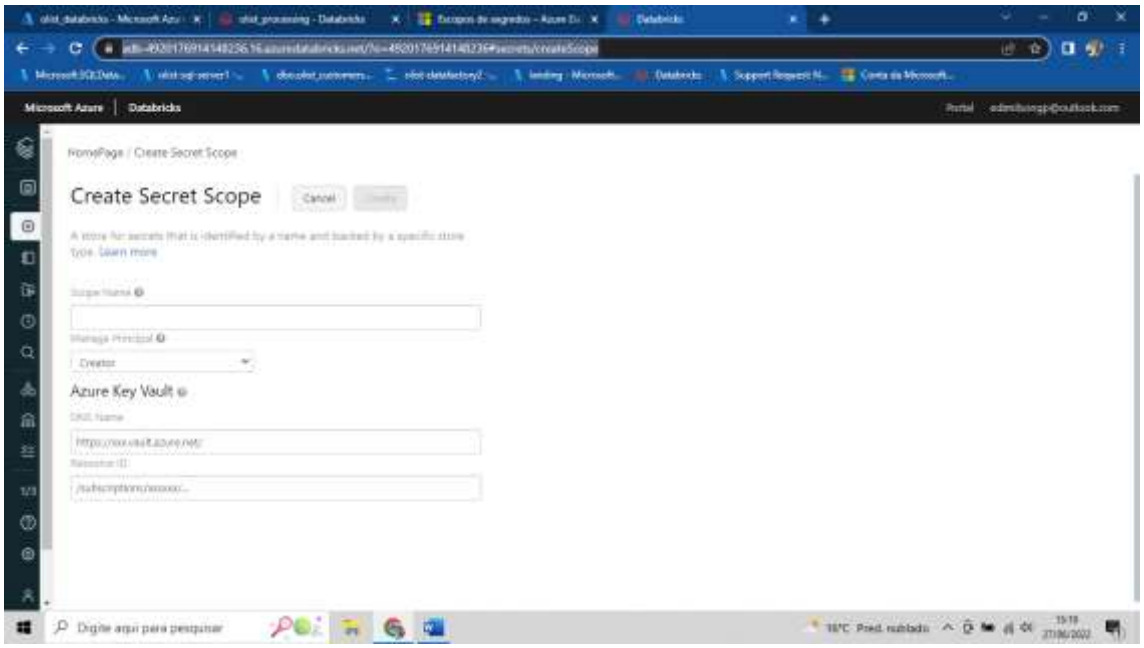
Copiando a URL

https://adb-4920176914148236.16.azuredatabricks.net



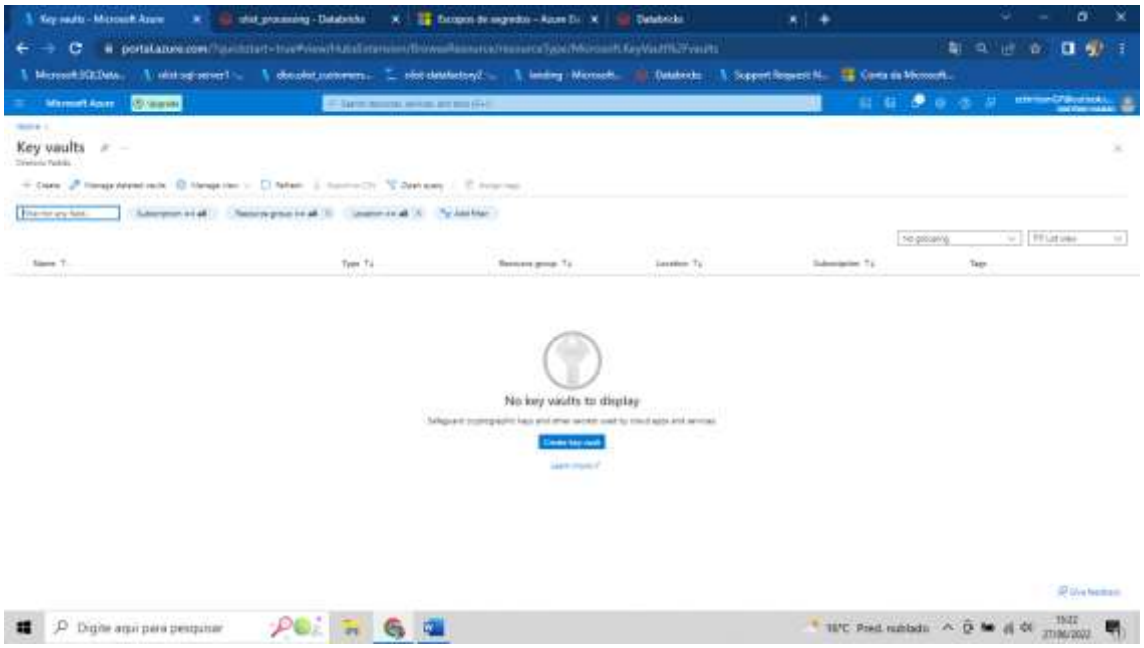
ALTERANDO A URL

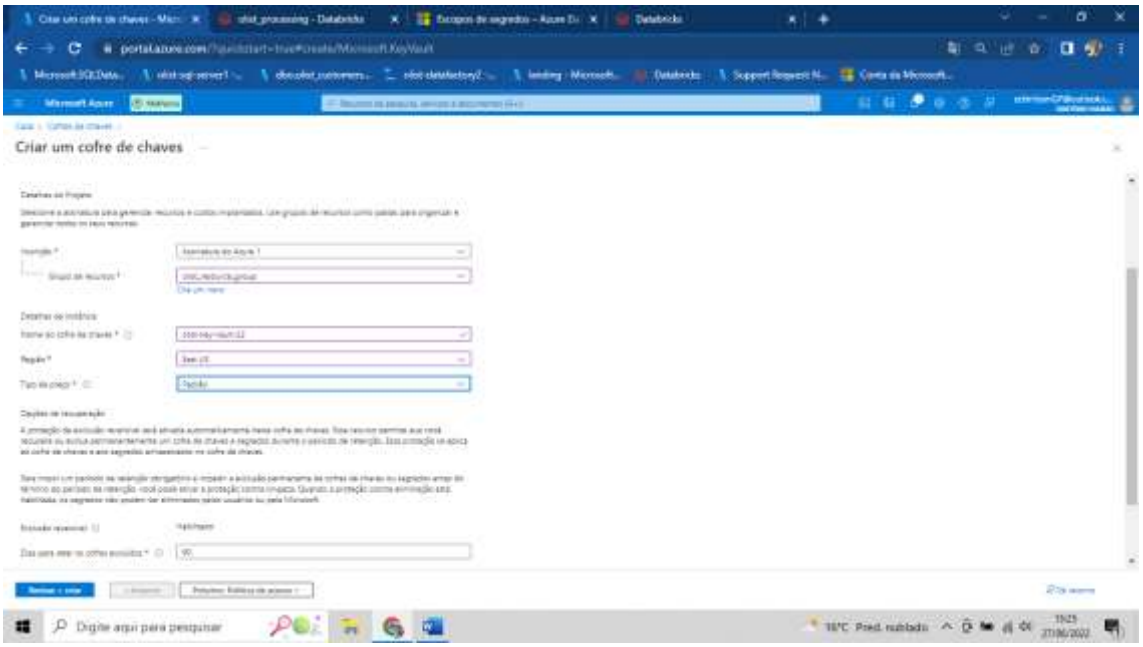
https://adb-4920176914148236.16.azuredatabricks.net/#secrets/createScope



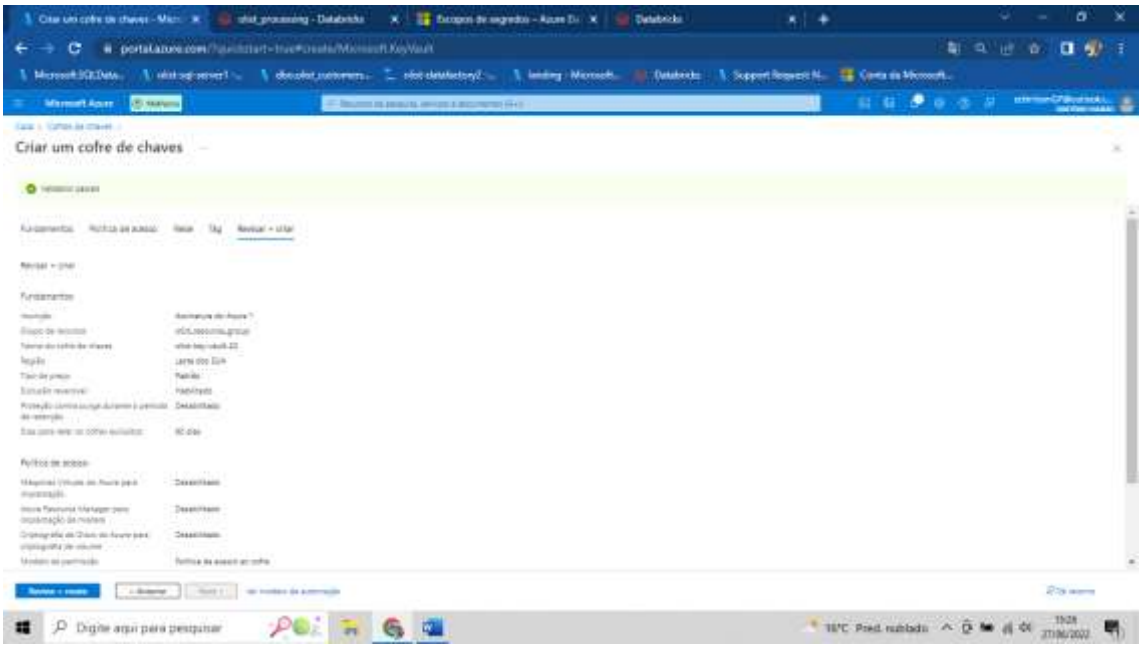
Criar AZURE KEY VAULT

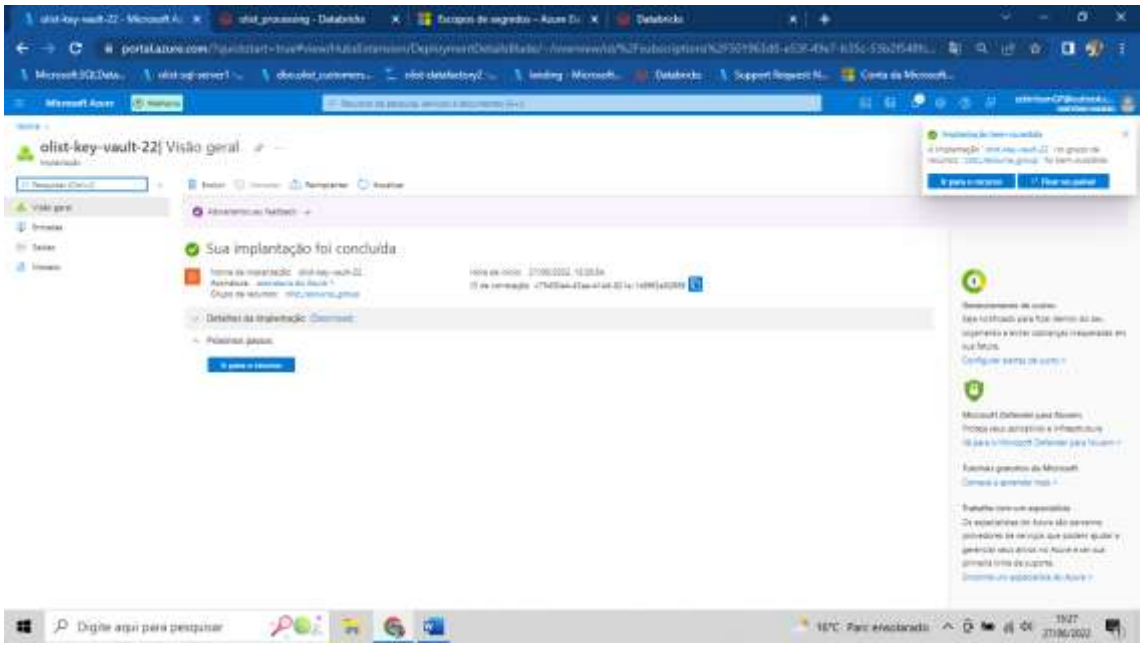
PESQUISAR PELO KEY VAULT



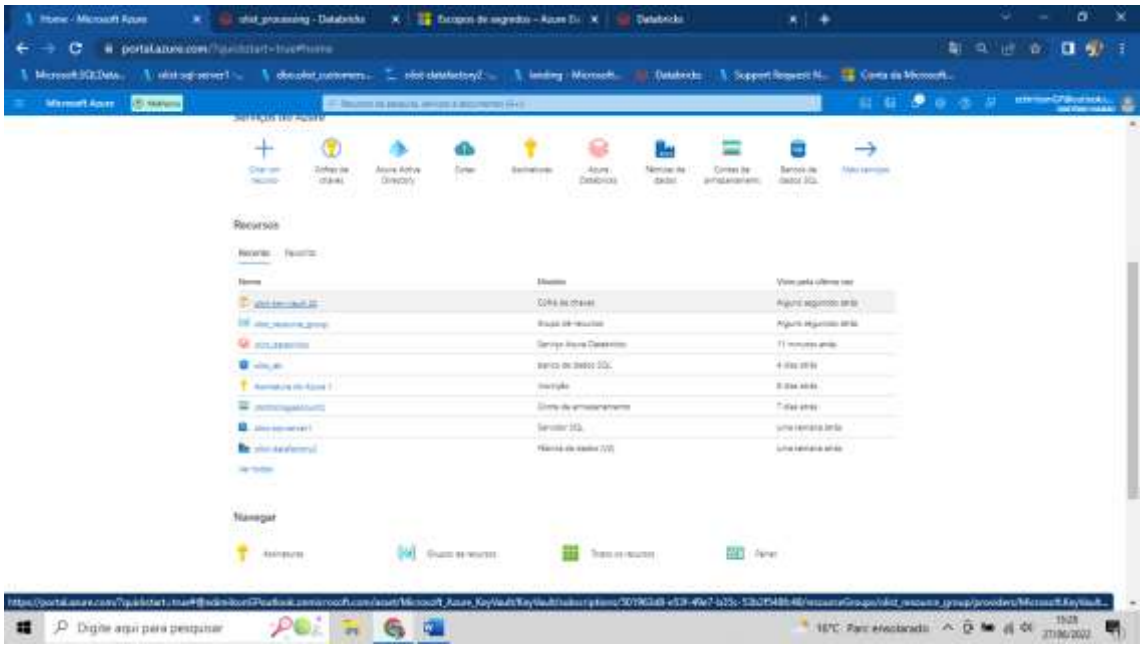


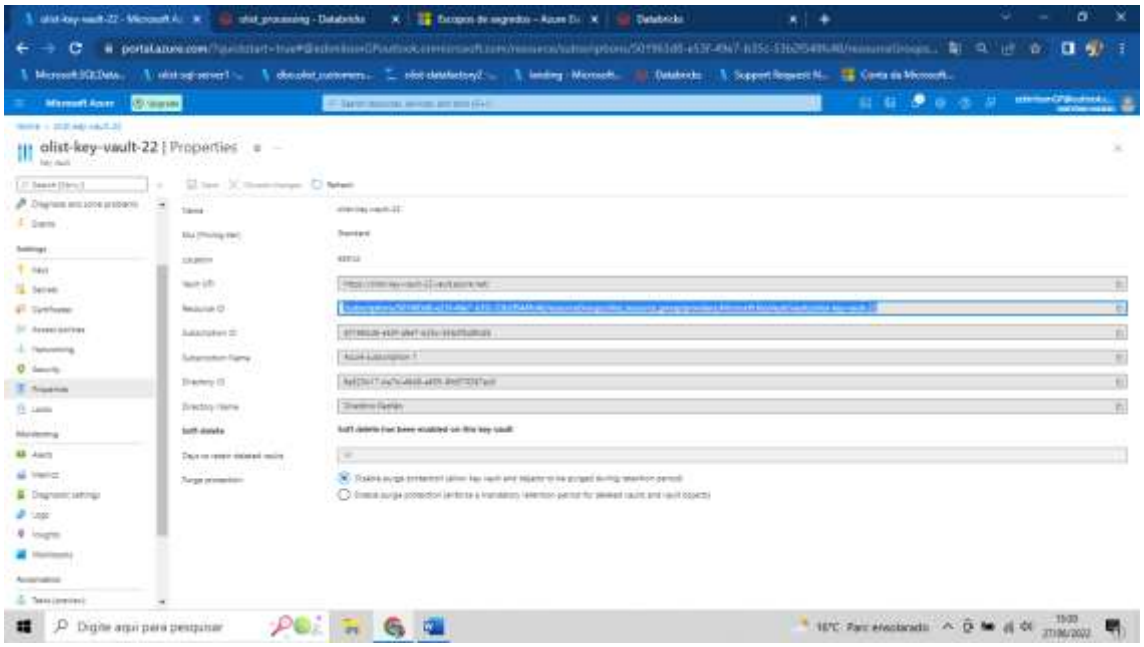
CLICAR EM PROXIMO ATÉ CRIAR E VALIDAR



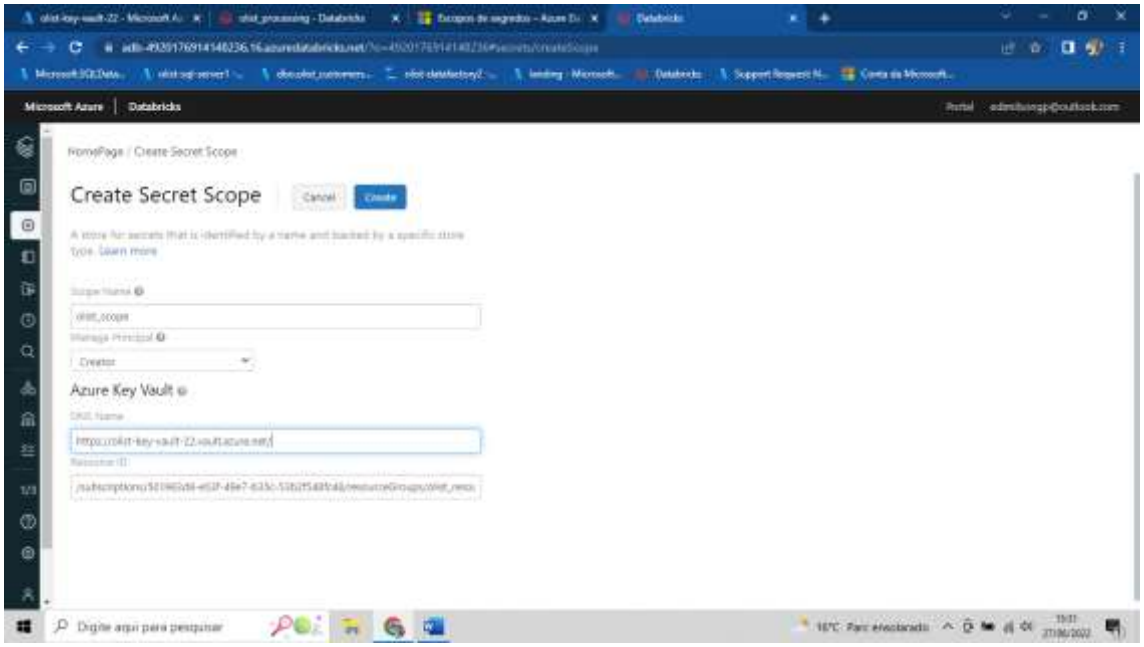


VOLTANDO NA PAGINA INICIAL JÁ IDENTIFICAMOS A CRIAÇÃO DA KEY VAULT-22

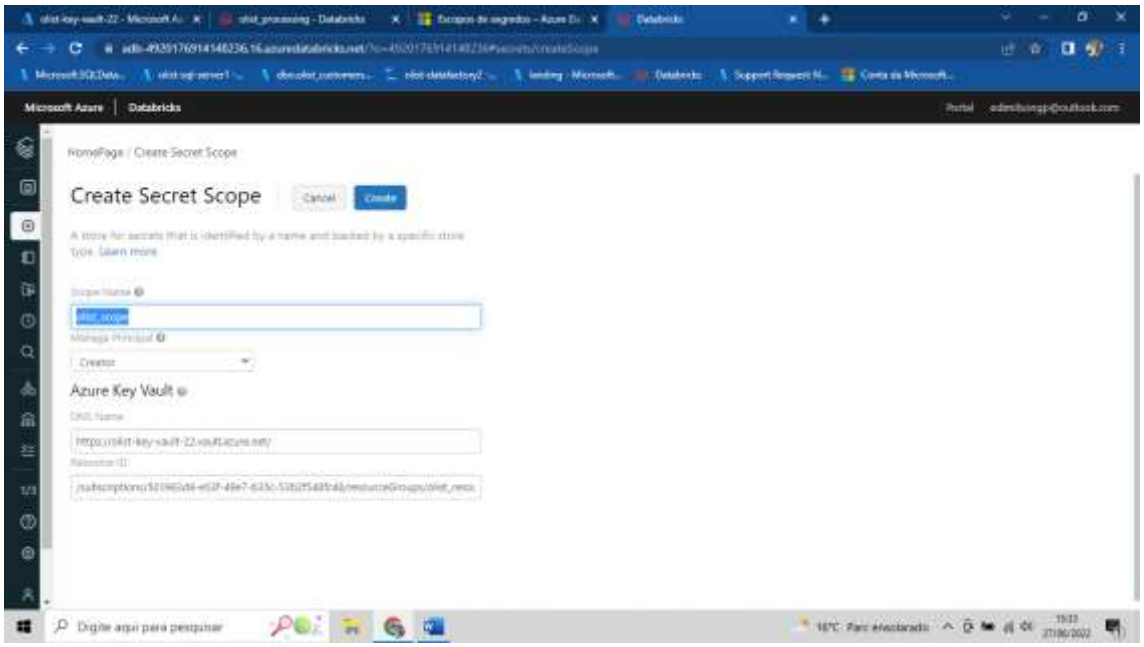




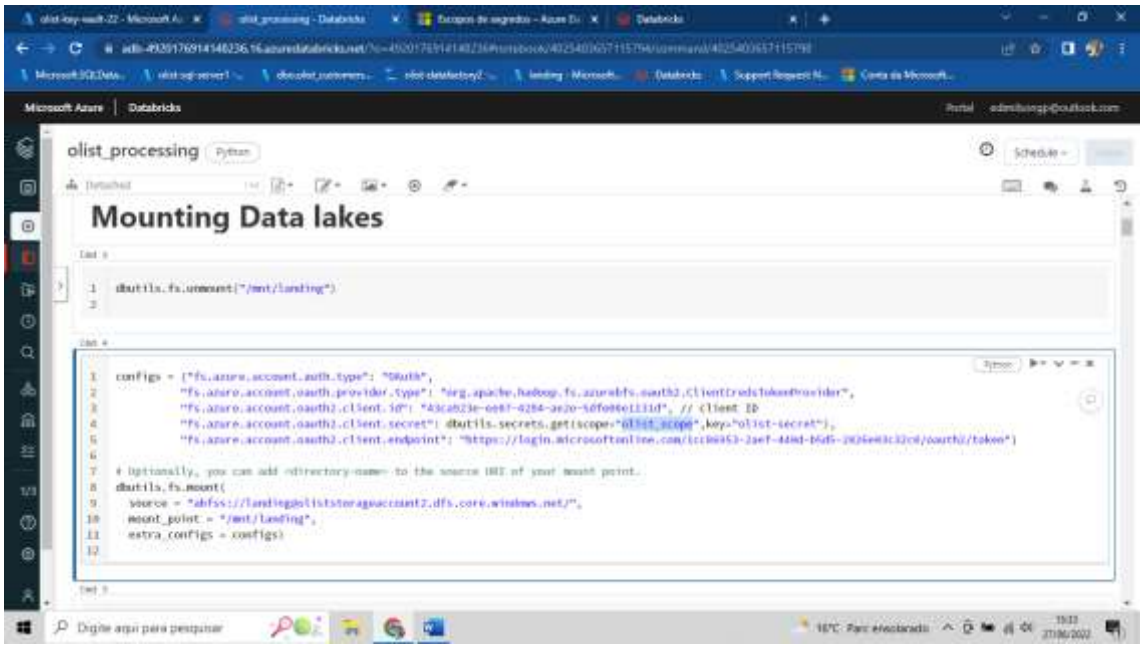
PEGANDO ESSE LINKS E COLOCANDO LÁ NA CRIAÇÃO DE SCOPO



COPIAR O NOME E CLICAR EM CRIAR

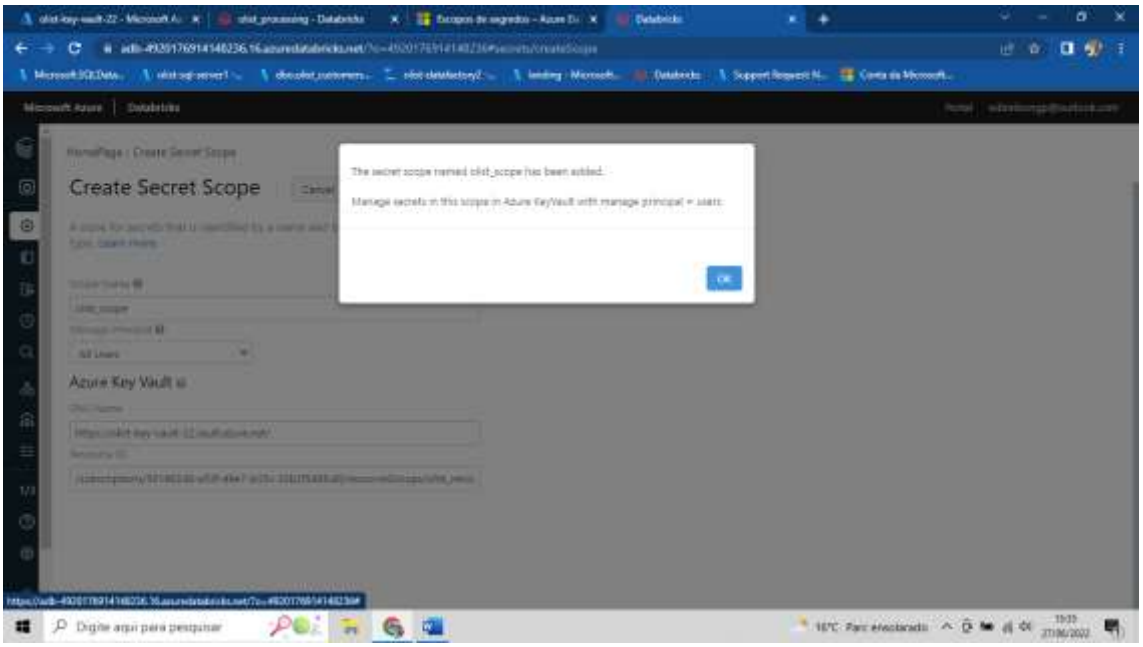


COLAR NO SCRIPT



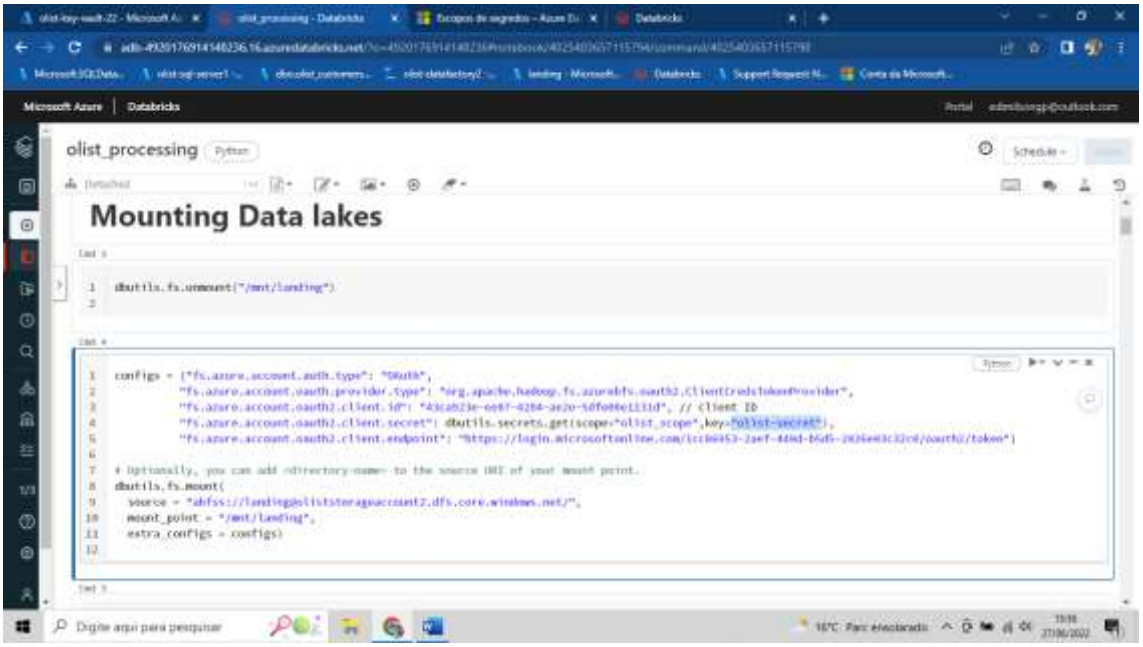
ALTERANDO PARA ALL USERS

CRIADO



AGORA

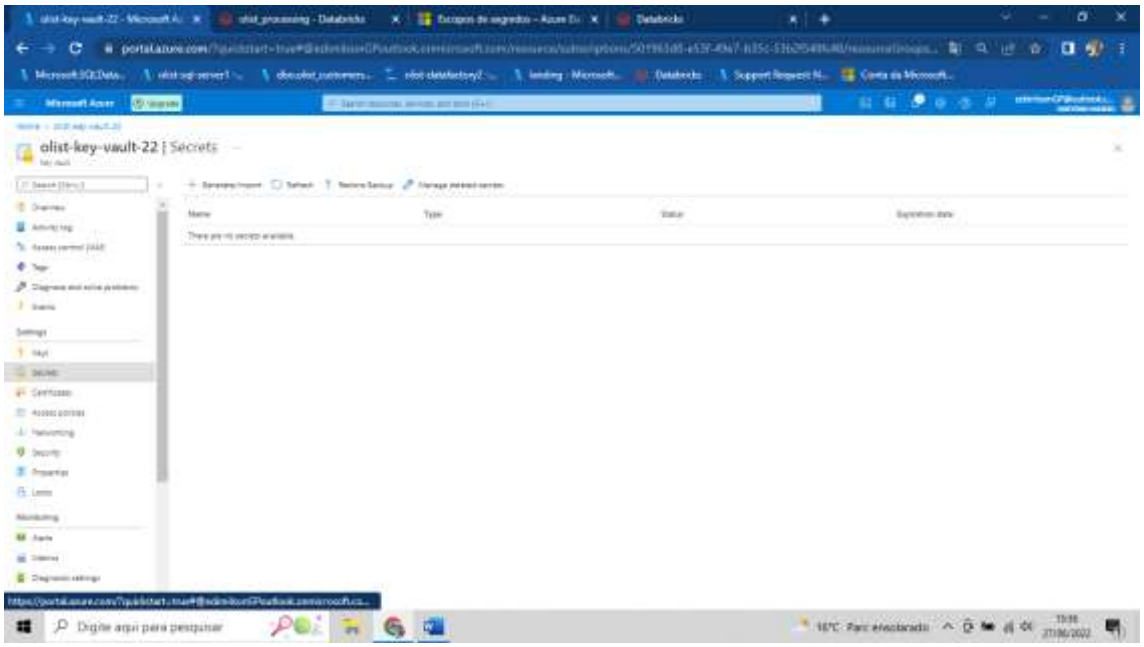
CRIAR O KEY



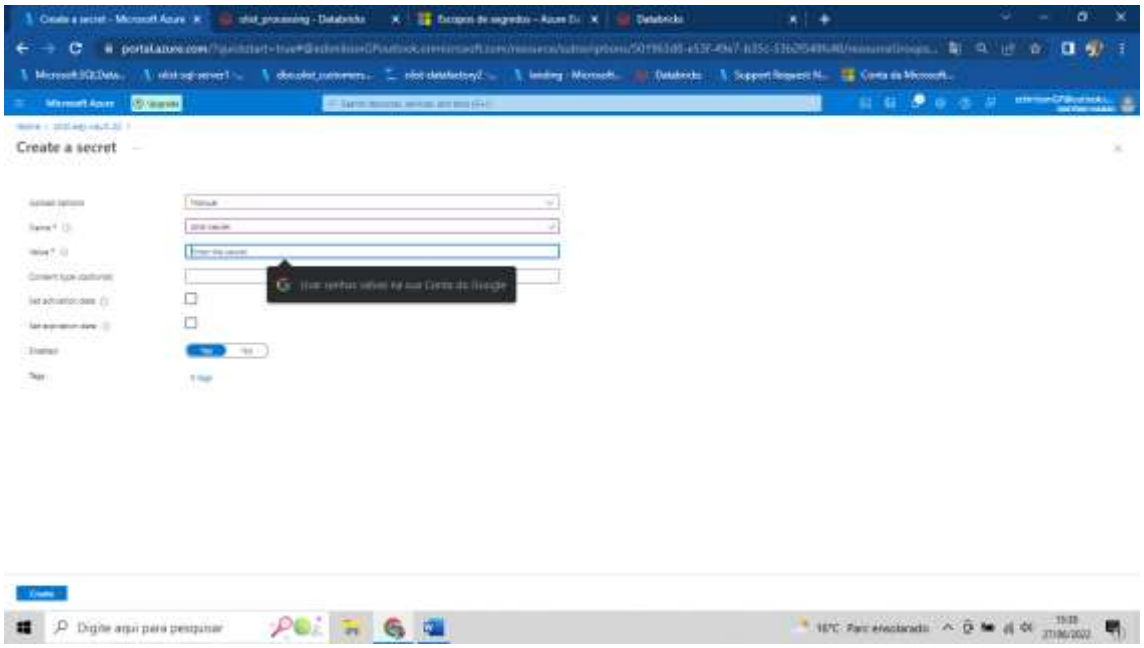
ENTRANDO NA KEY VAULT

EM SECRETS

CLICAR EM GERAR E IMPORTAR



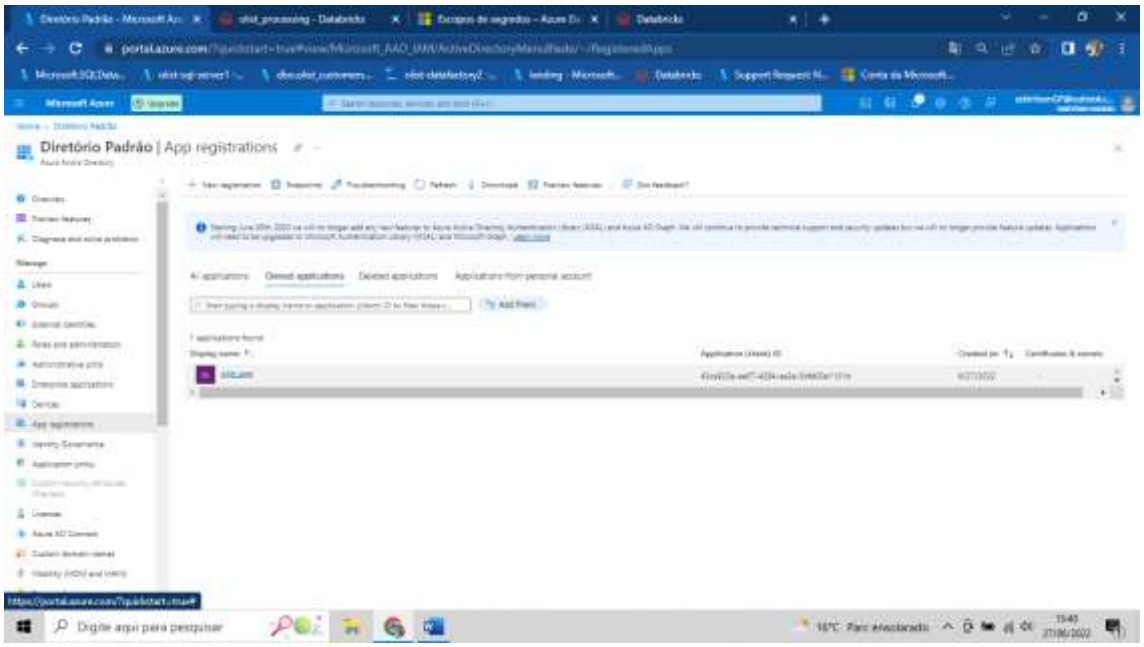
Criando o SECREDO



IR PARA AZURE ACTIVE

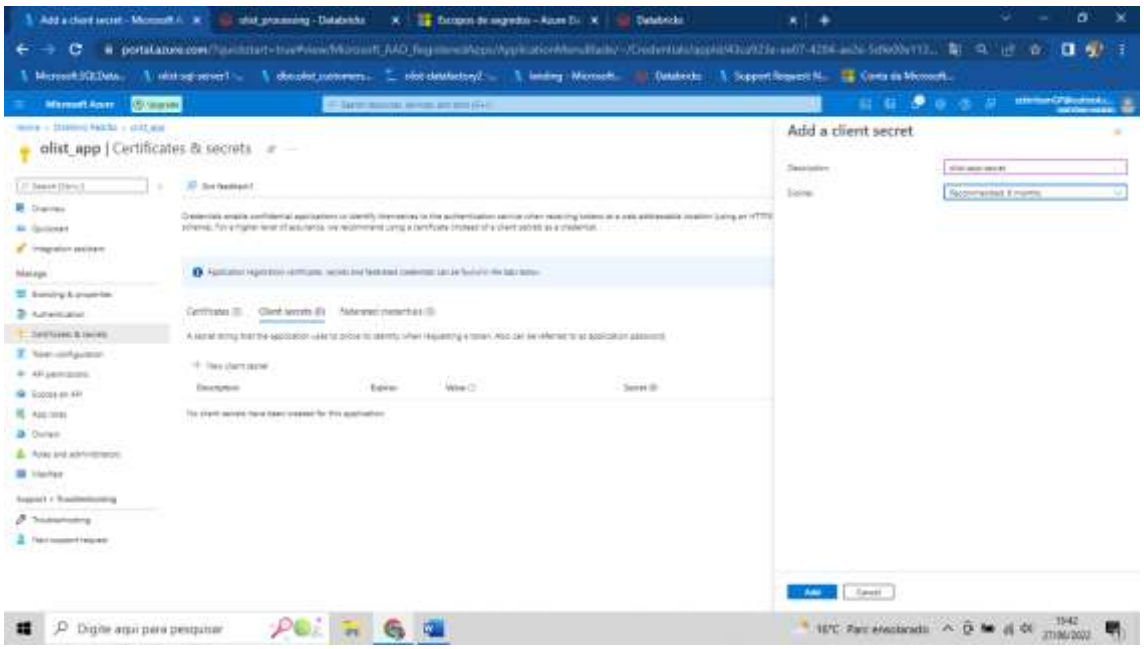
IR EM REGISTRO DE APLICATIVO

CLICAR NO APLICATIVO

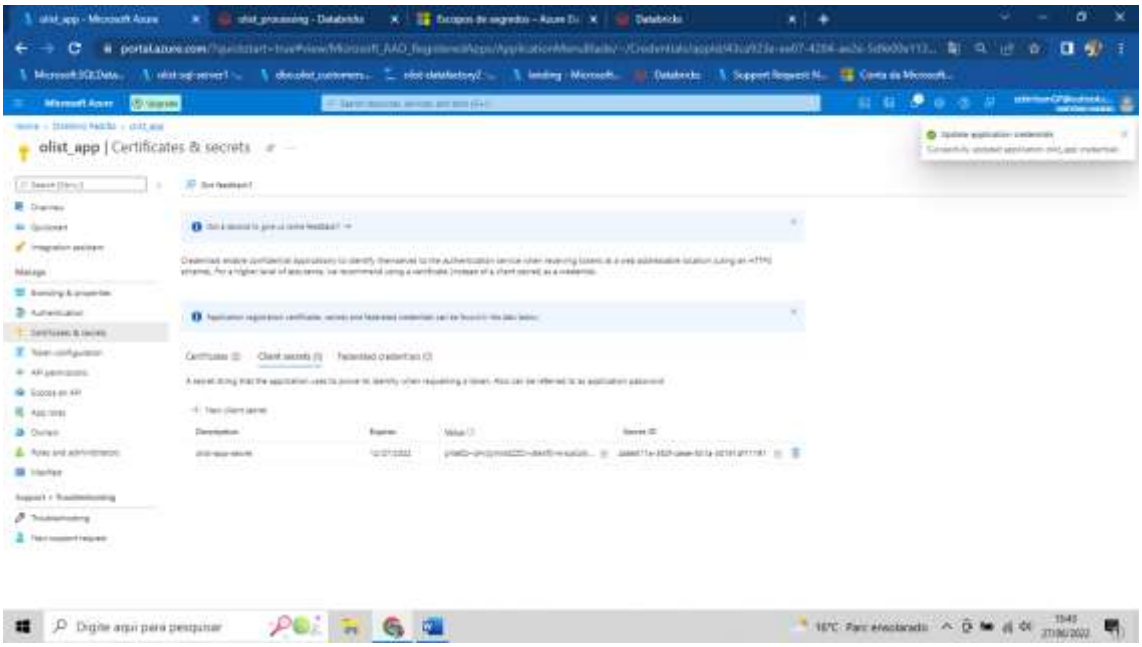


Clicar em novo segredo do cliente e preencher as informações

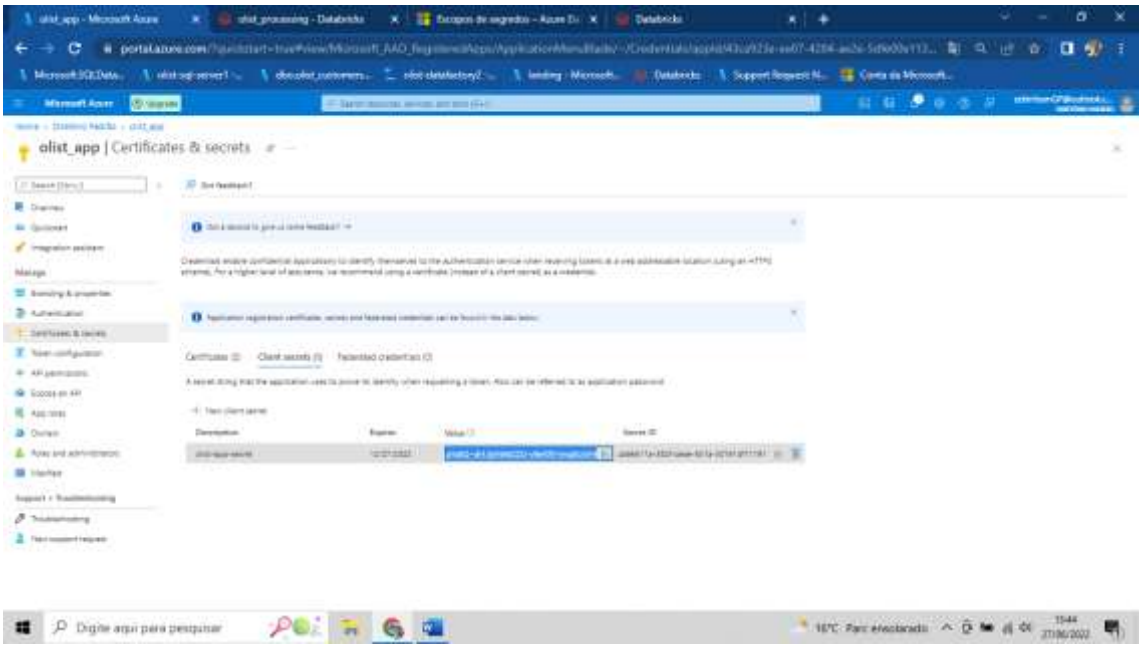
CLICAR EM ADD



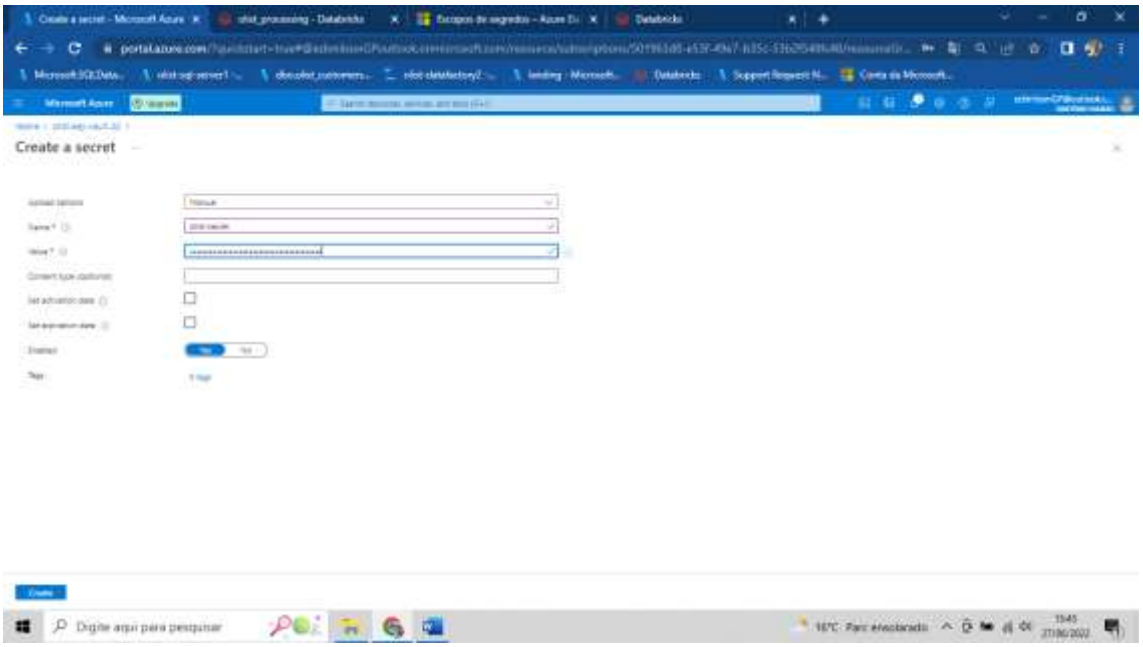
SEGREDO CRIADO



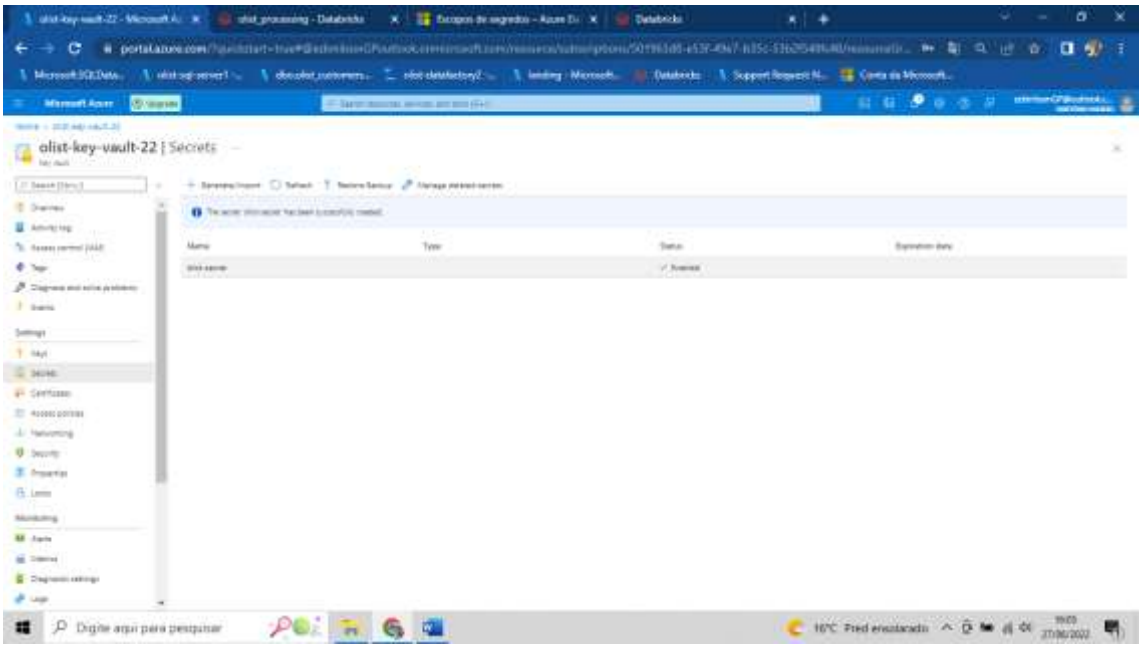
COPIA pYa8Q~dH.0yhWdZZO-vBeKf0-hrsqKo5iHhzLaKN



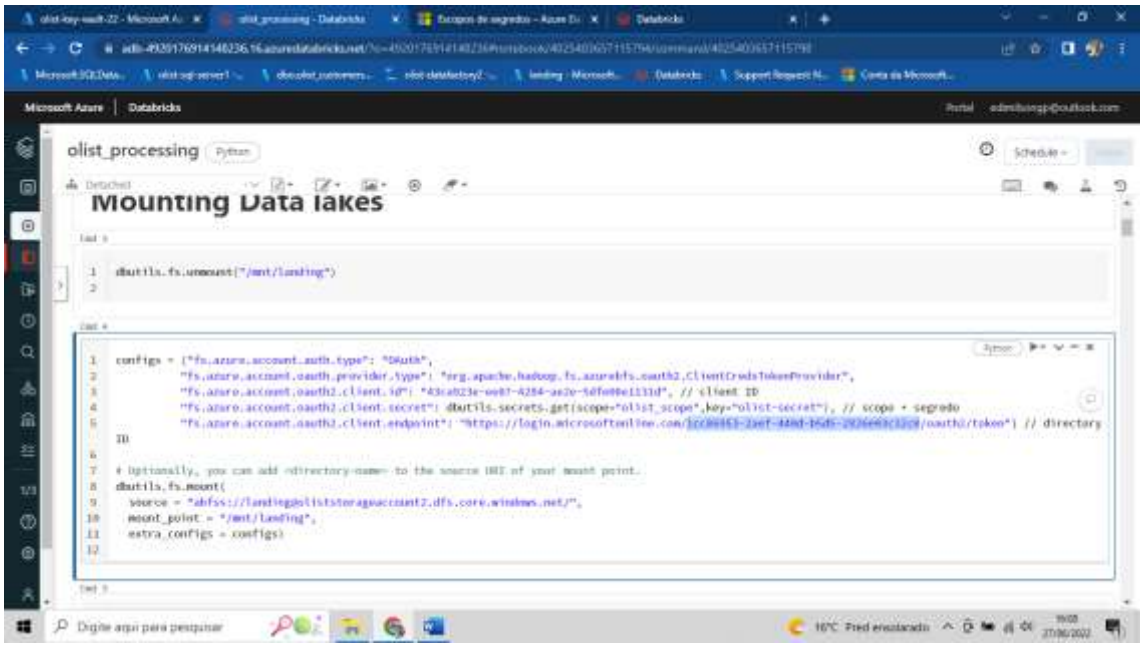
IR EM KEY VAULT



CRIADO O OLIST SECRET

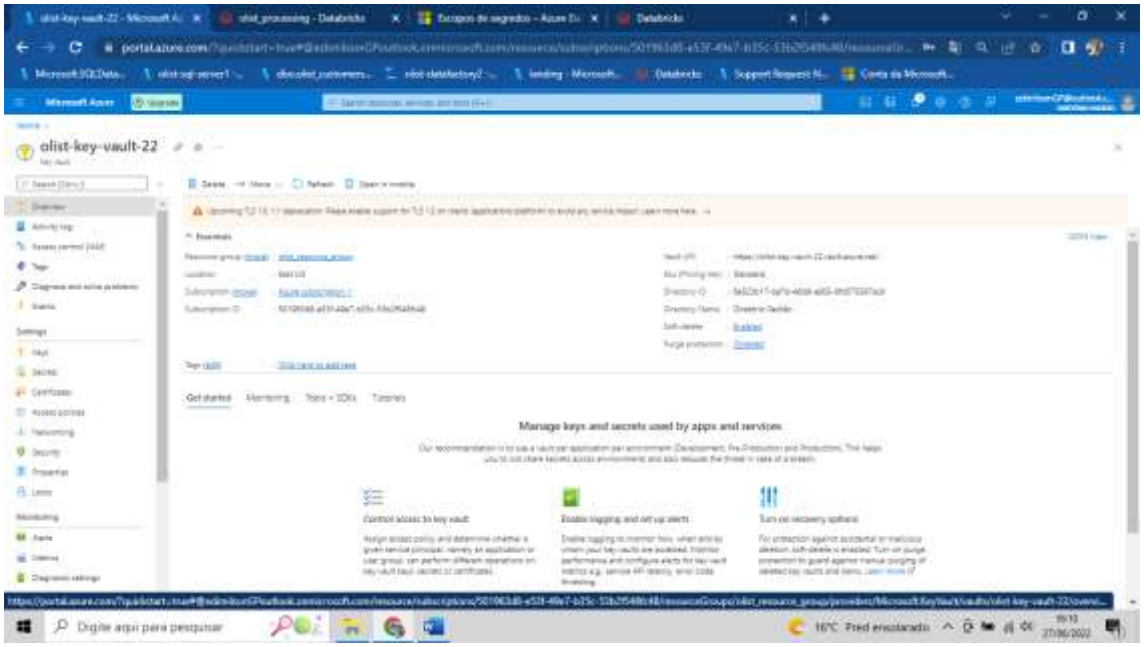


SUBSTITUIR



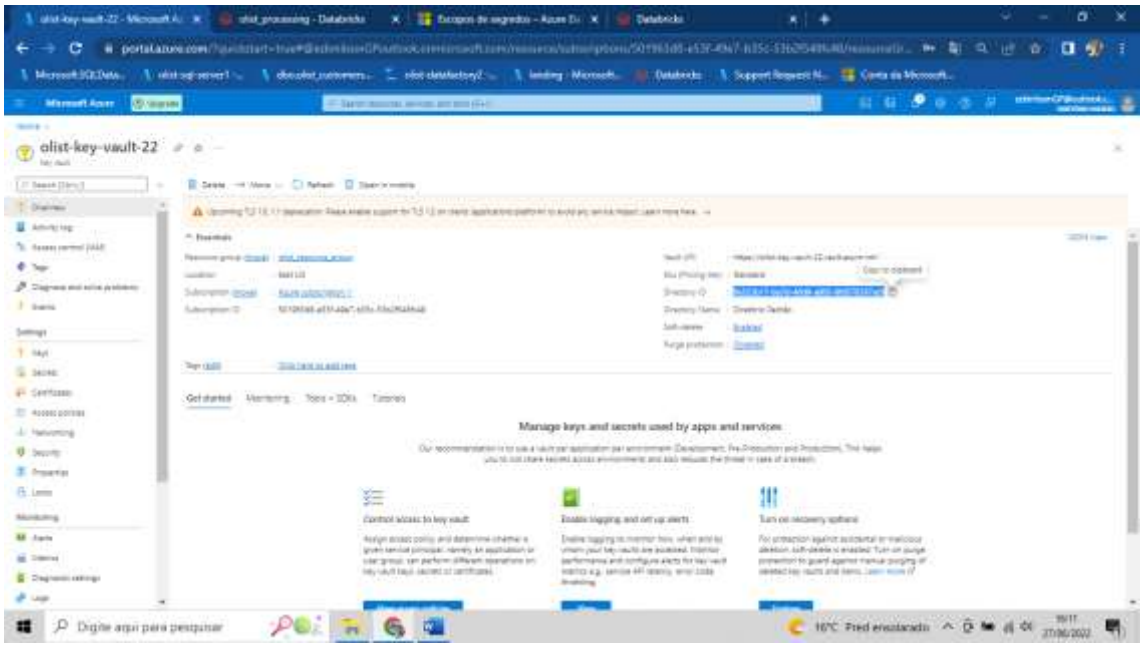
PARA PEGAR ESSAS INFORMAÇÕES DEVEMOS IR EM KEY VAULT

VISÃO GERAL

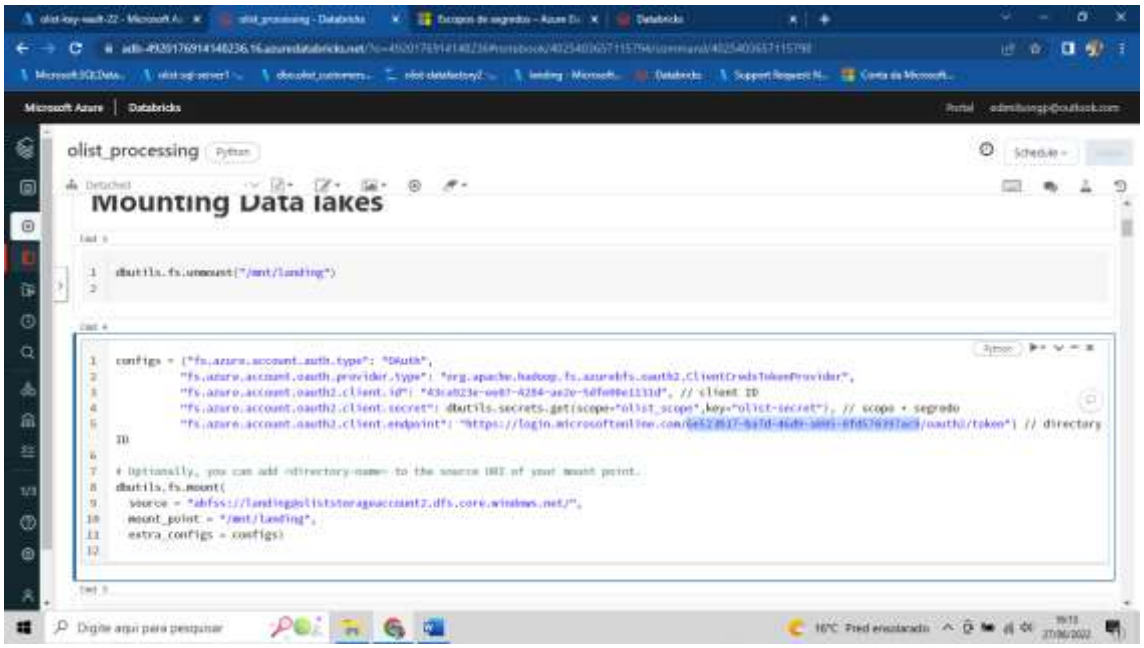


ID DO DIRETORIO

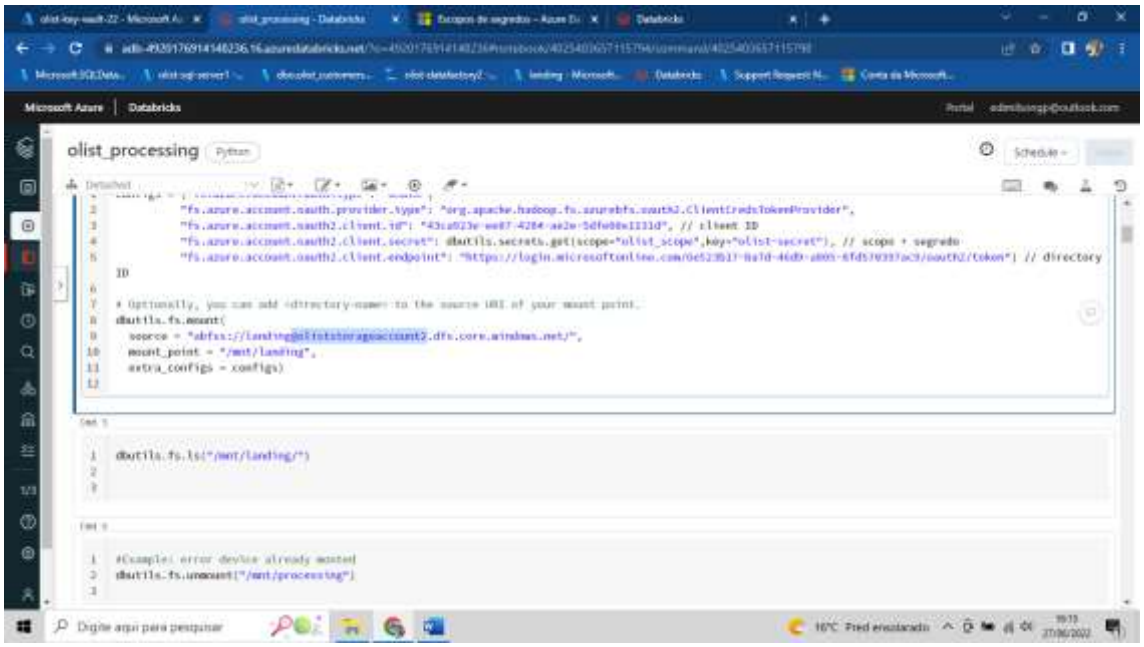
6e523b17-ba7d-46d9-a805-8fd570397ac9



SUBSTITUINDO



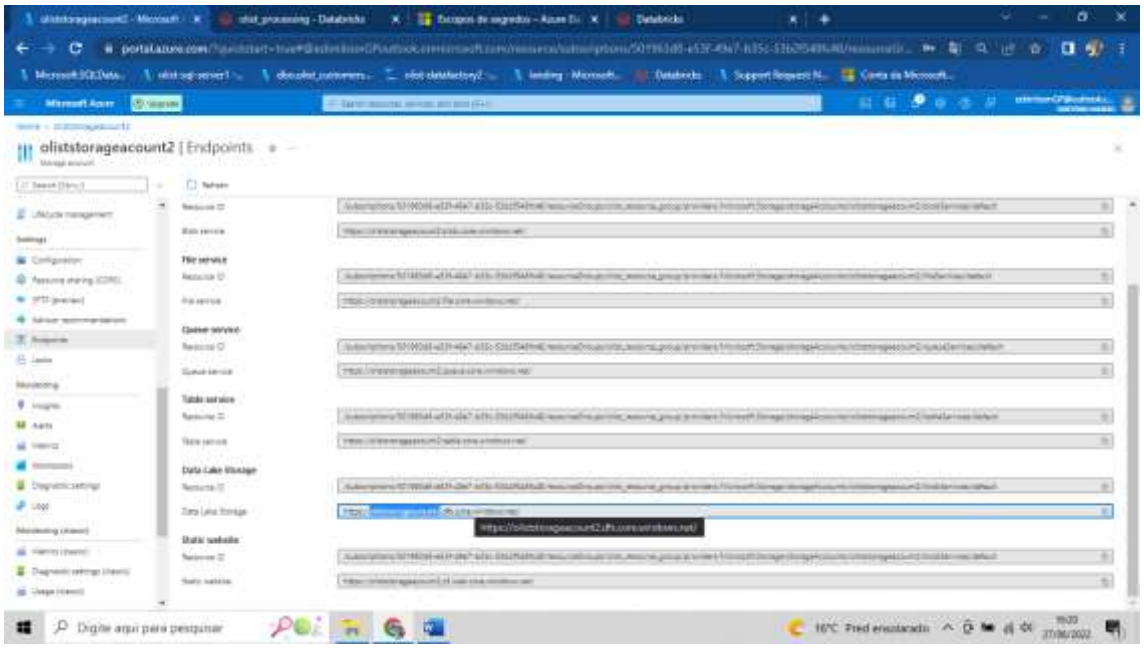
AGORA VAMOS ALTERAR A URL DA STORAGE ACOUNT



IR PARA PAGINA INICIAL

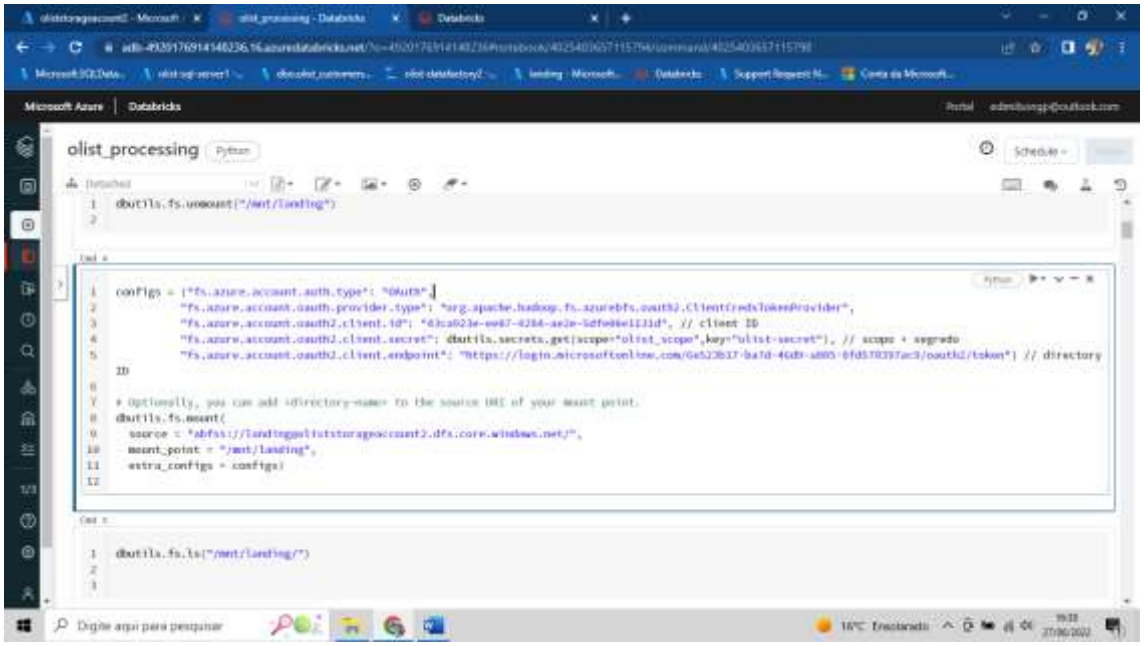
E EM STORAGE ACOUNT

COPIAR O CODIGO oliststorageacount2



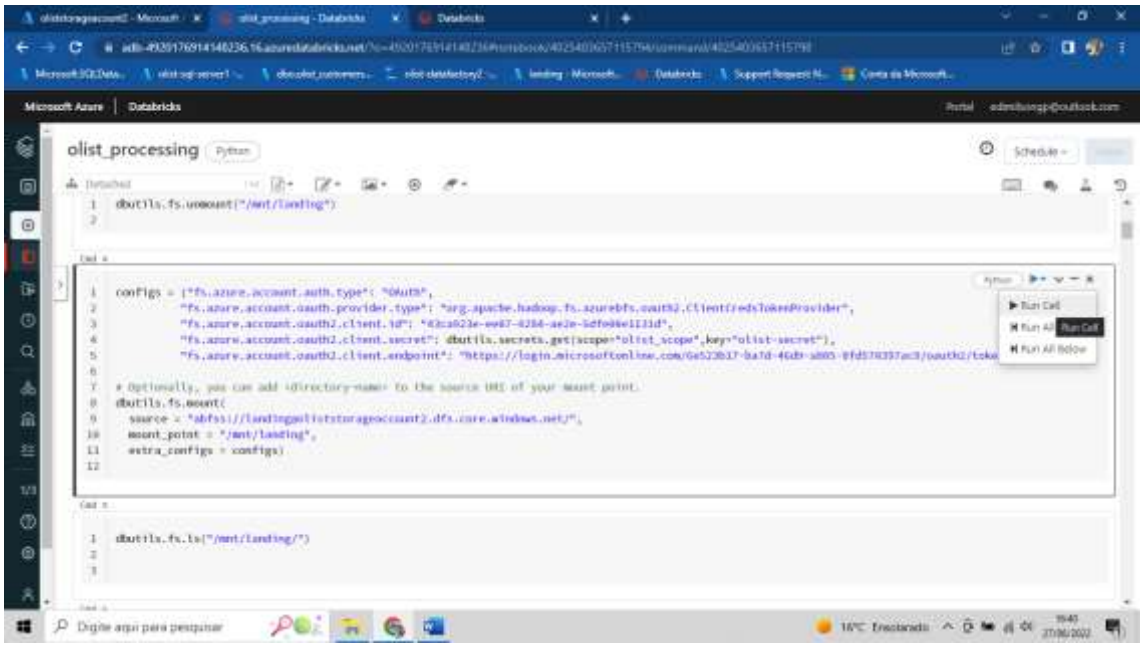
COLAR

E PRONTO MONTAGEM FINALIZADA

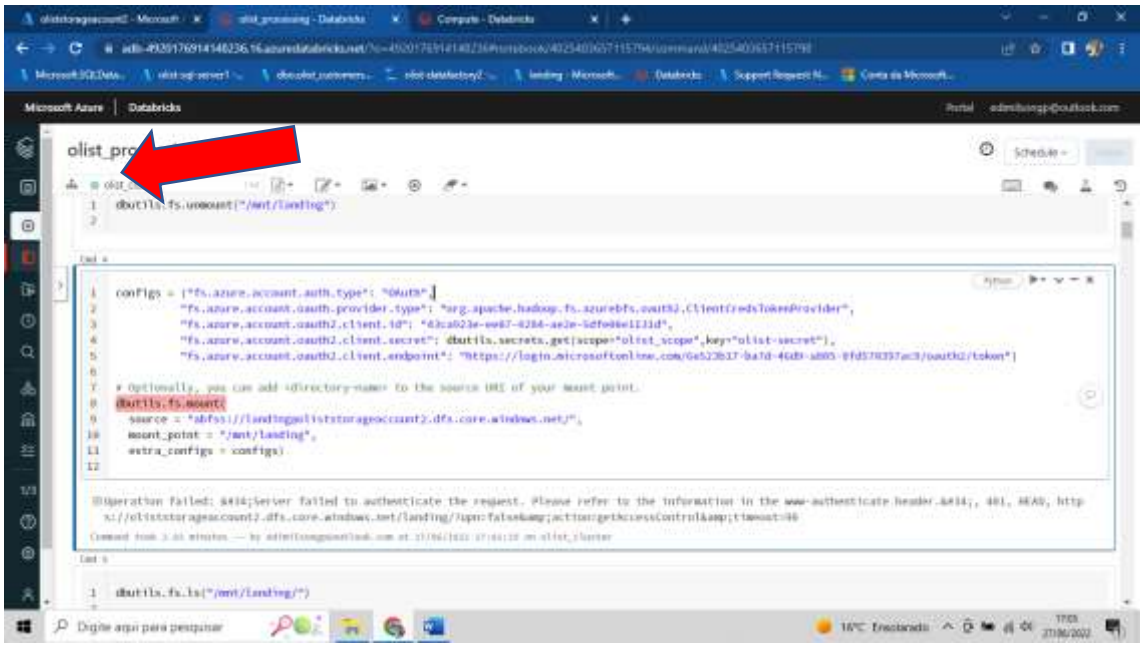


EXECUTAR O SCRIPT

CLICANDO EM RUN CELL

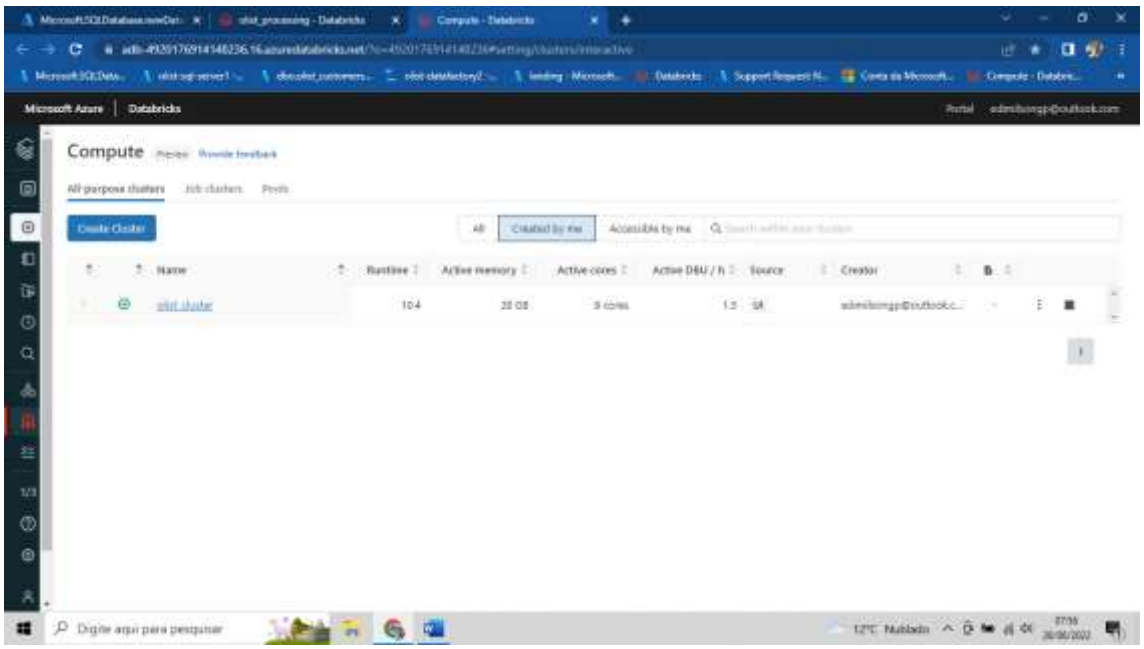


EM EXECUÇÃO



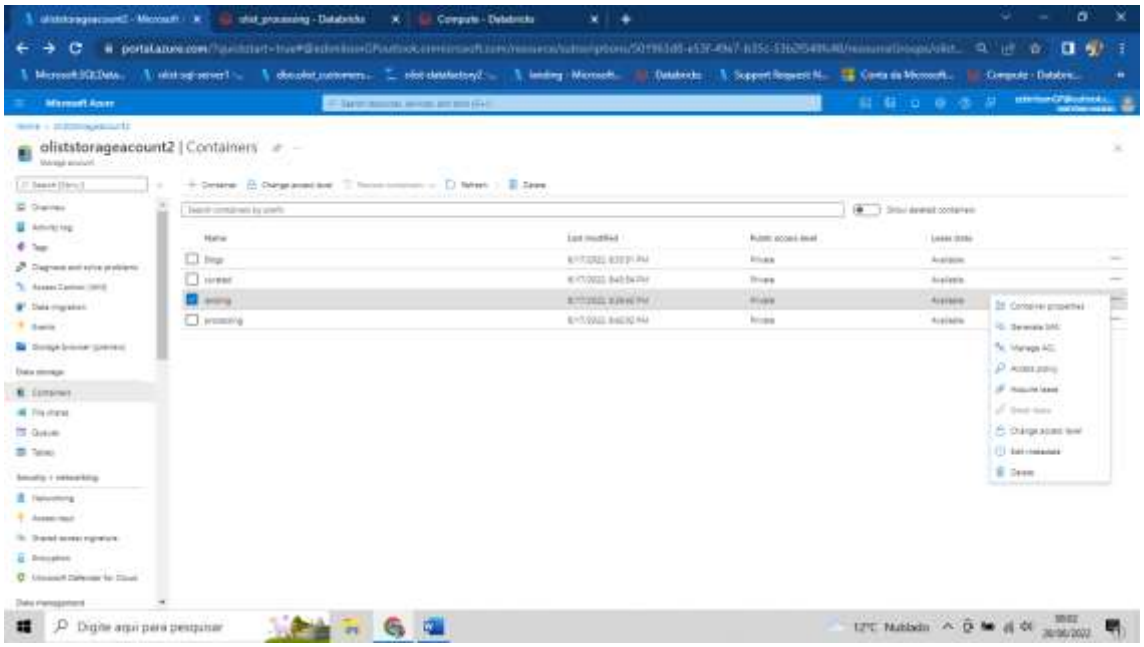
VERIFICANDO SE O CLUSTER NO DATABRICKS ESTÁ ATIVO

ESTÁ RODANDO



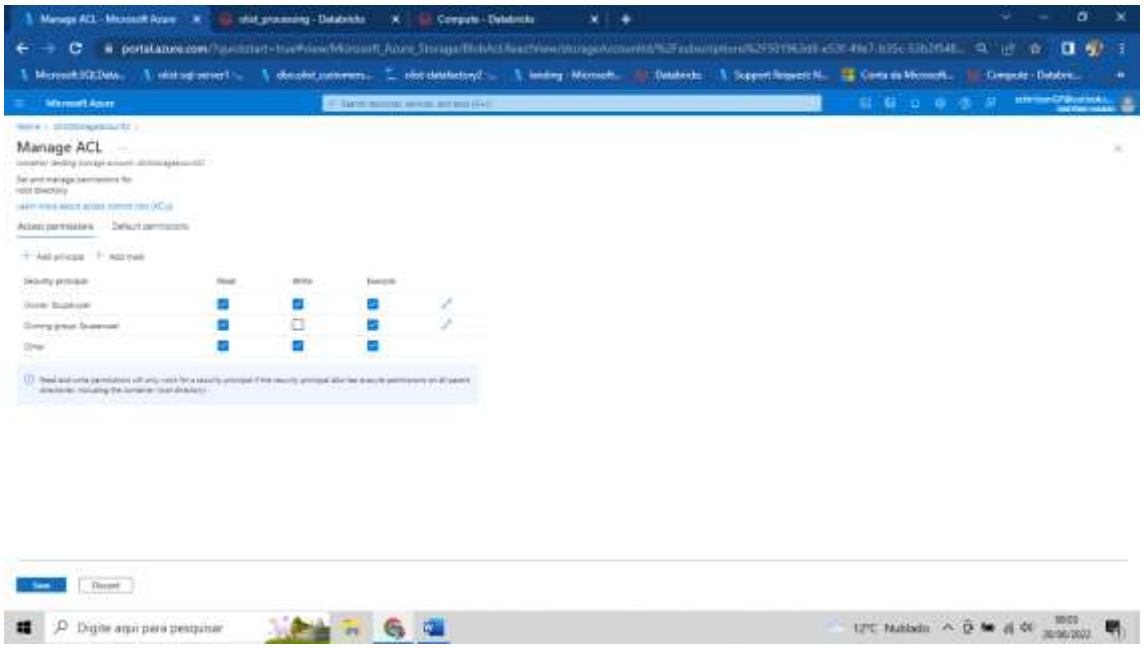
AGORA VOU LIBERAR AS PERMIÇÕES

DIRETAMENTE NO CONTEINER



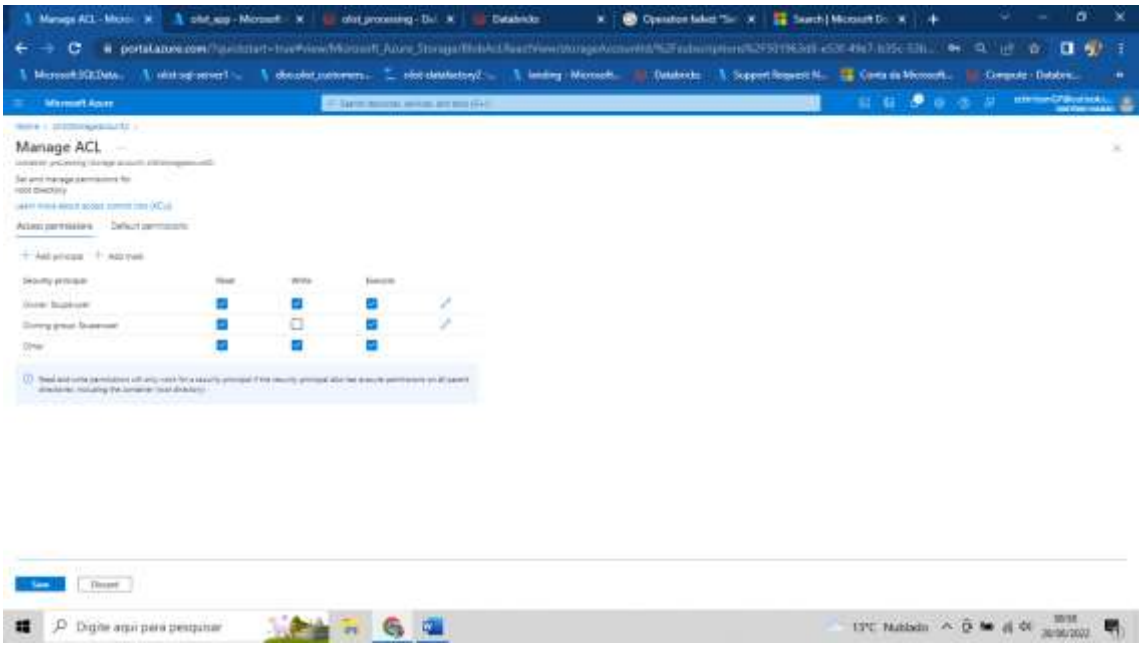
PERMIÇÕESATIVADAS PARA OUTROS USUARIOS

SALVAR

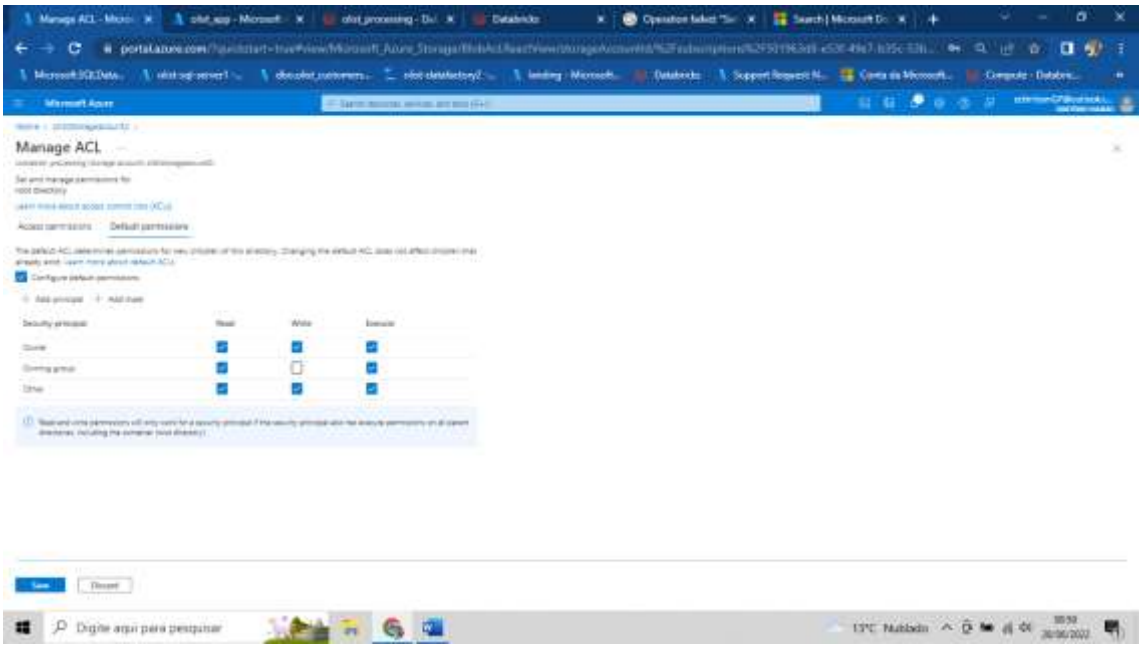


Agora fazer o mesmo processo de permissão para os demais conteiners

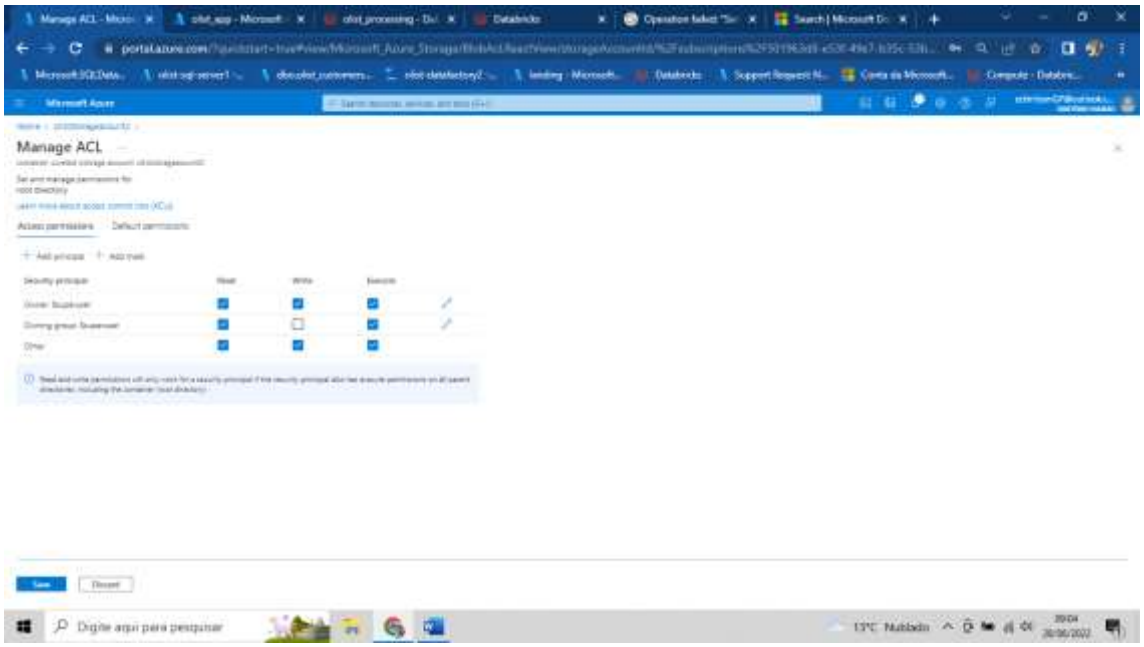
PROCESSING

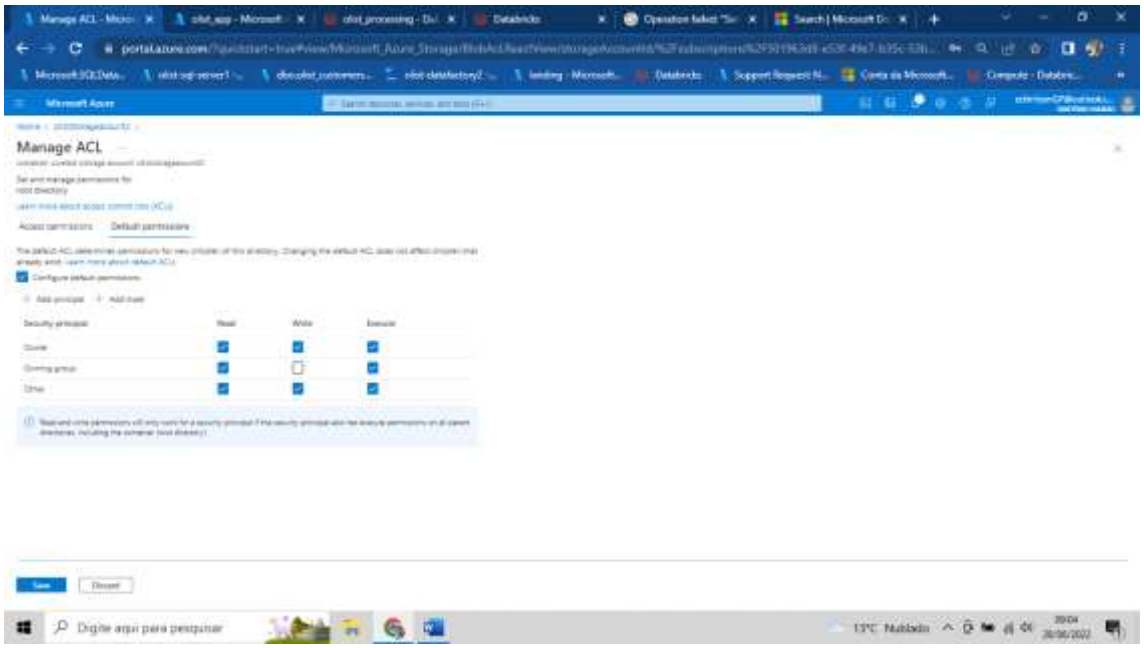


Também em permissão padrão



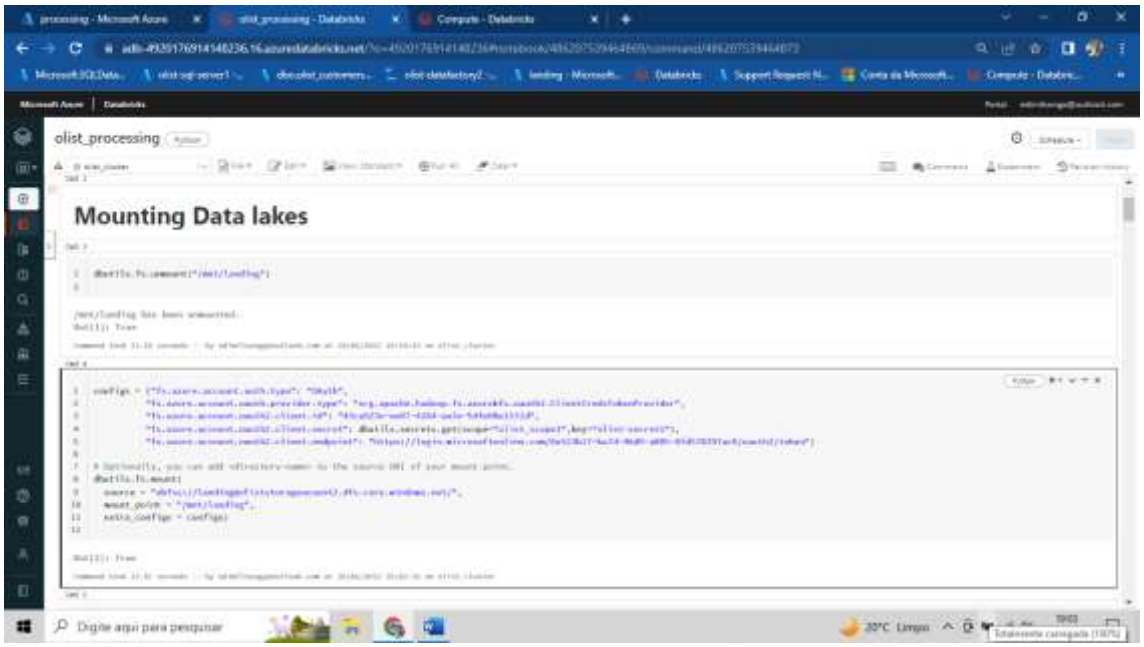
CURATED

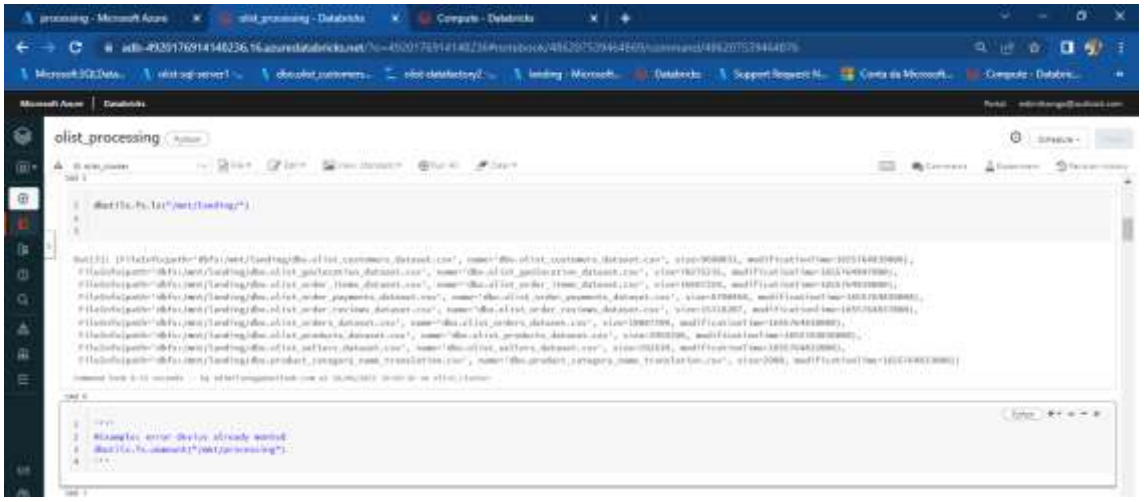




AGORA SÓ RODAR O NOTEBOOK EXECUTANDO UM POR UM,

CASO DE ERRO RODAR O NOTEBOOK DO INICIO

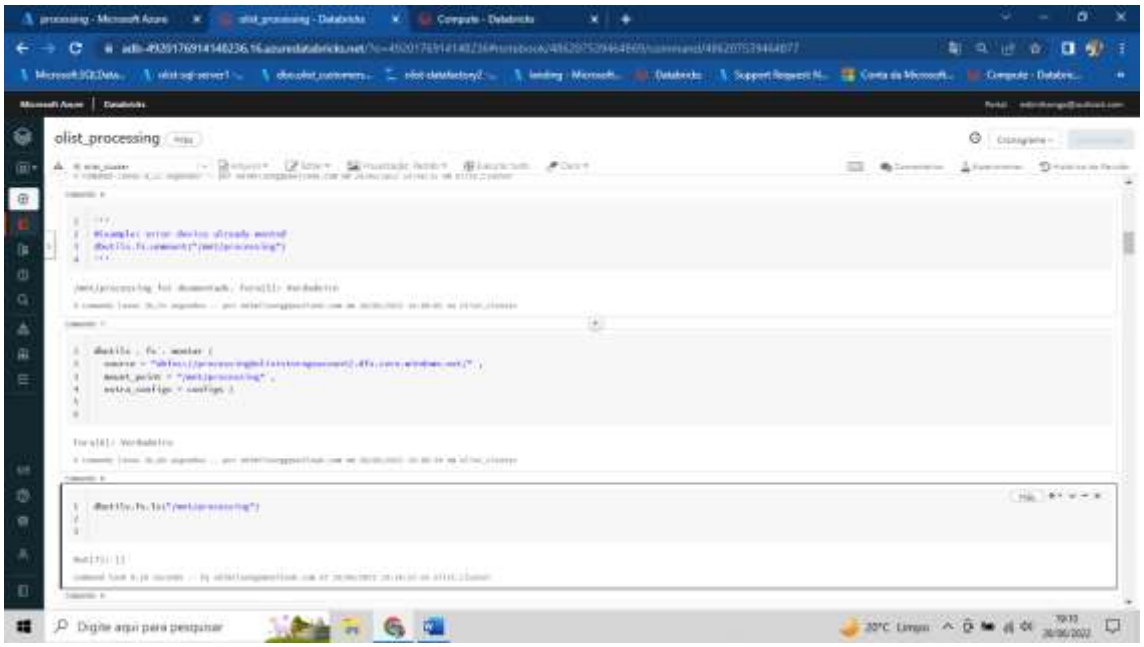




CASO APRESENTE A INFROMAÇÃO DE PROCESSING JÁ CRIADO EXECUTAR O

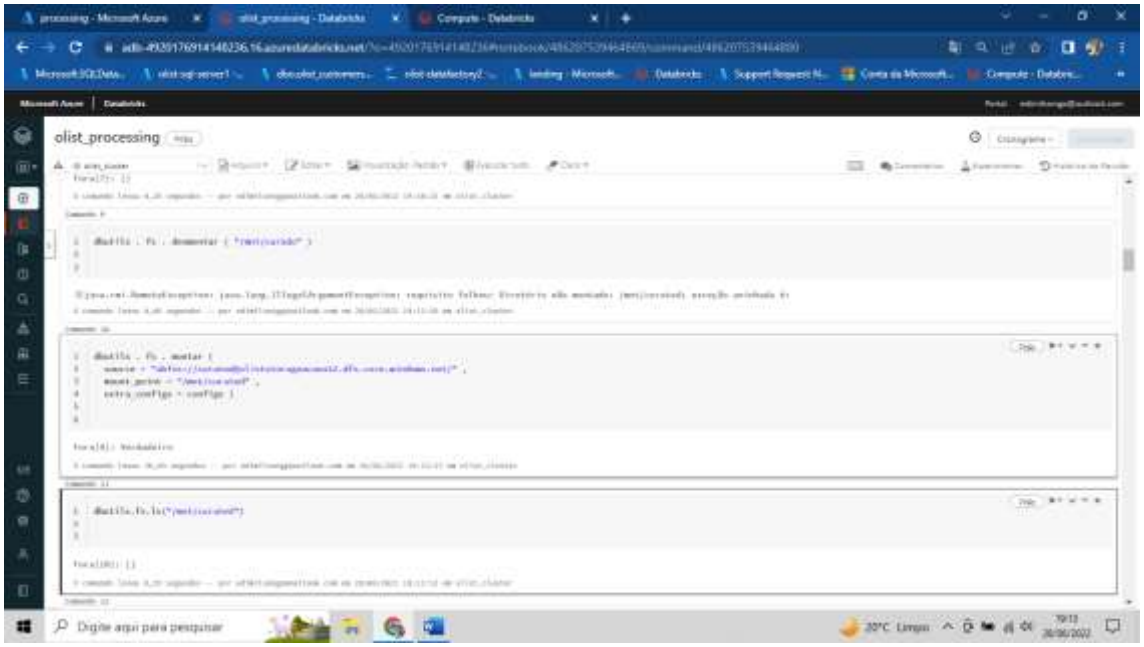
#Example: error device already monted

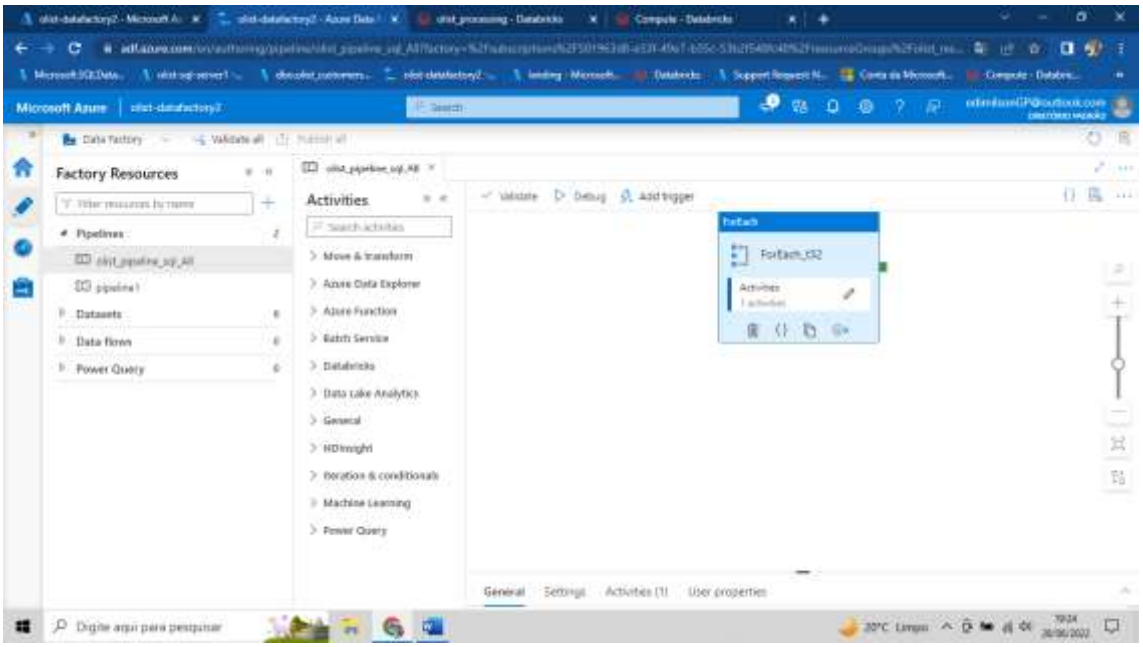
dbutils.fs.unmount("/mnt/processing") FORA DAS ‘ ‘ ASPAS PARA DEMONTAR



FAZER A MESMA COISA CASO CURATED ESTEJA CRIADO

EXECUTADOS ATÉ AQUI

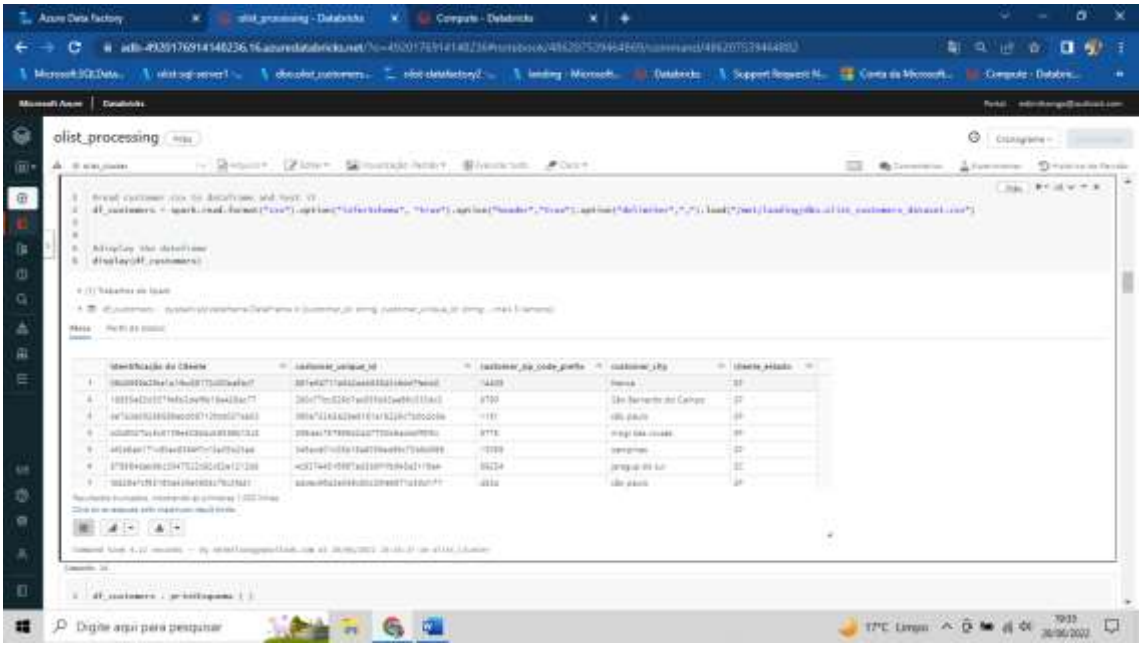




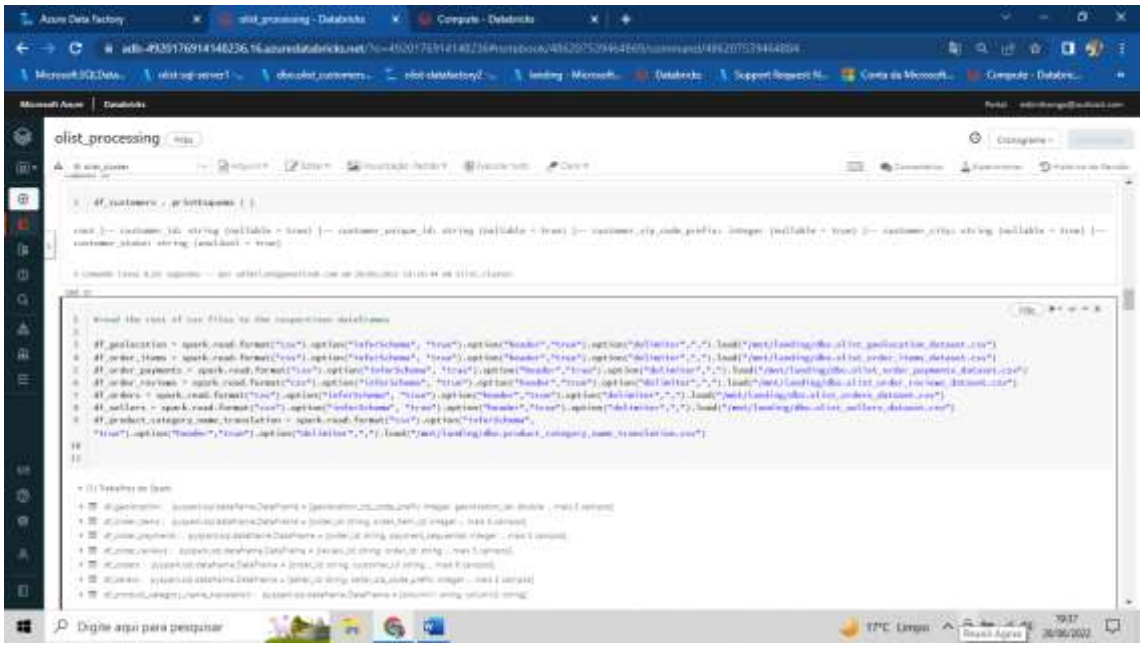
ORIGEM

E DISPARAR GATILHO

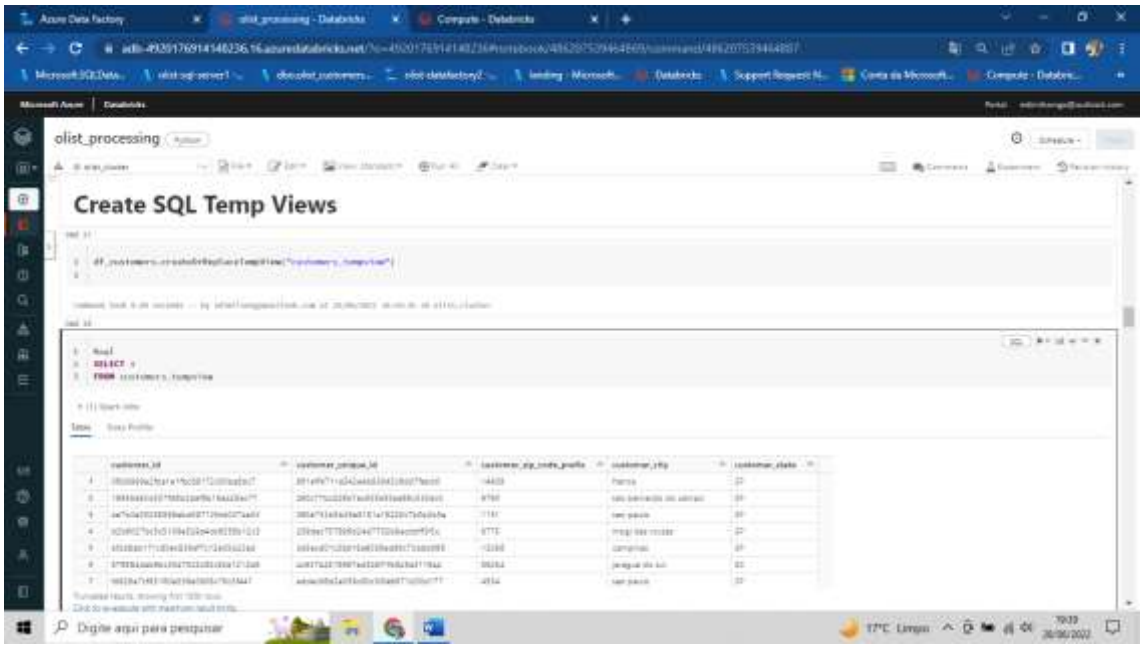
EXECUTANDO A LEITURA



EXUTANDO AS OUTRAS CELULAS



CRIANDO SQL



PODEMOS DROPAR UMA AS TABELA UMA POR UMA OU FAZER

DE ACORDO COM A TABELA QUE JÁ EXISTE

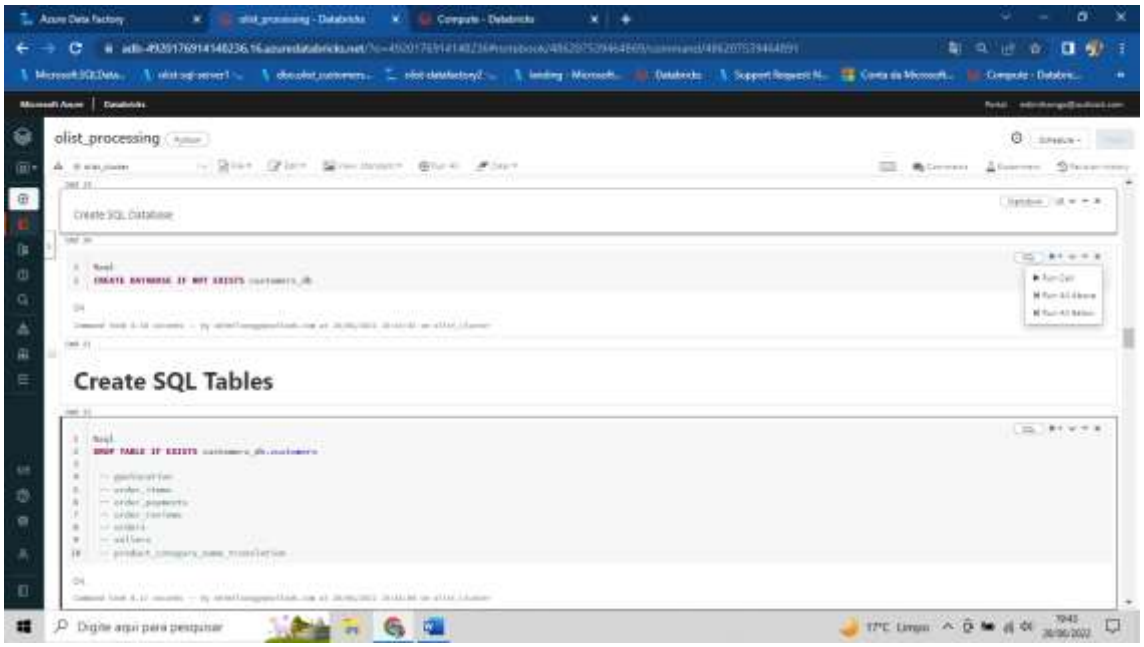
%sql

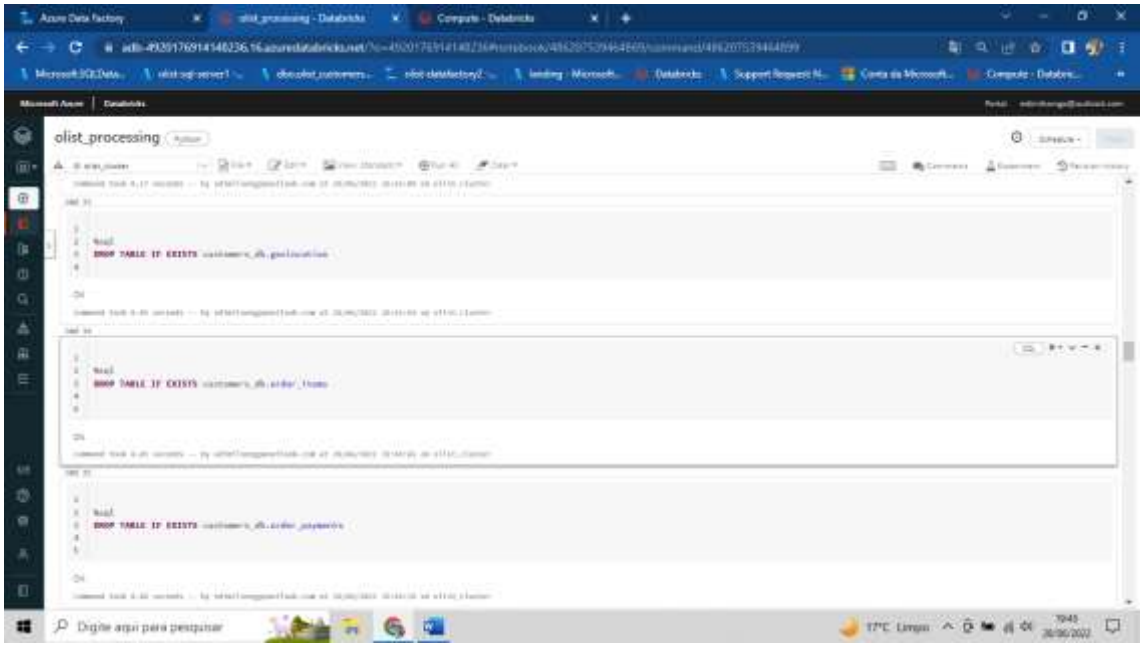
CREATE TABLE IF NOT EXISTS customers_db.customers

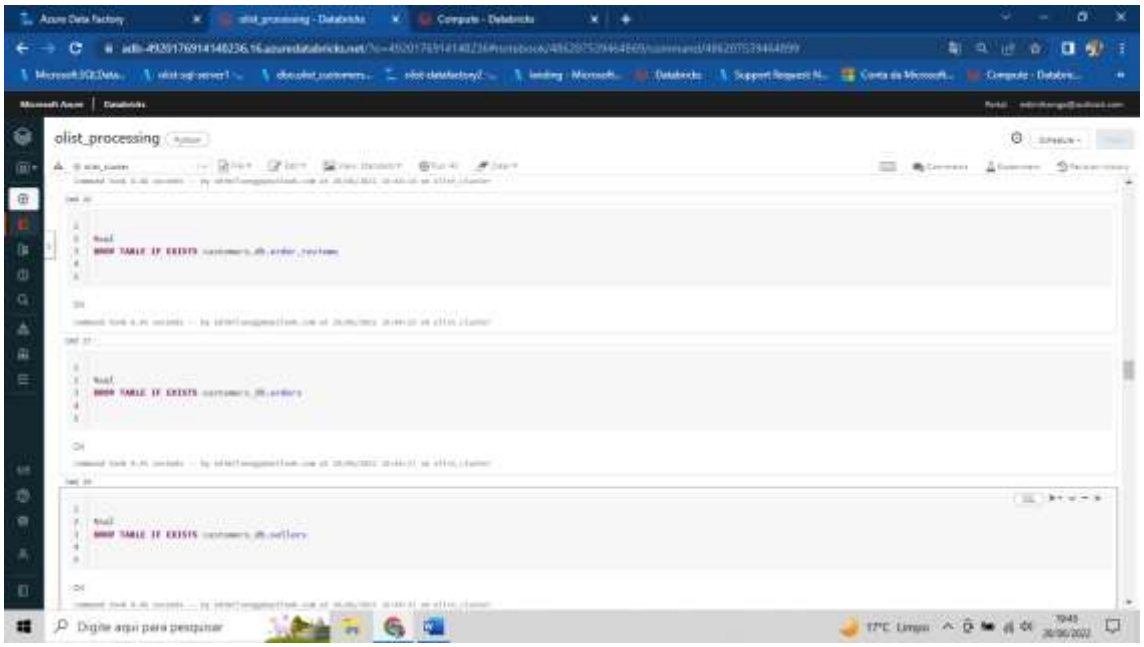
USING CSV

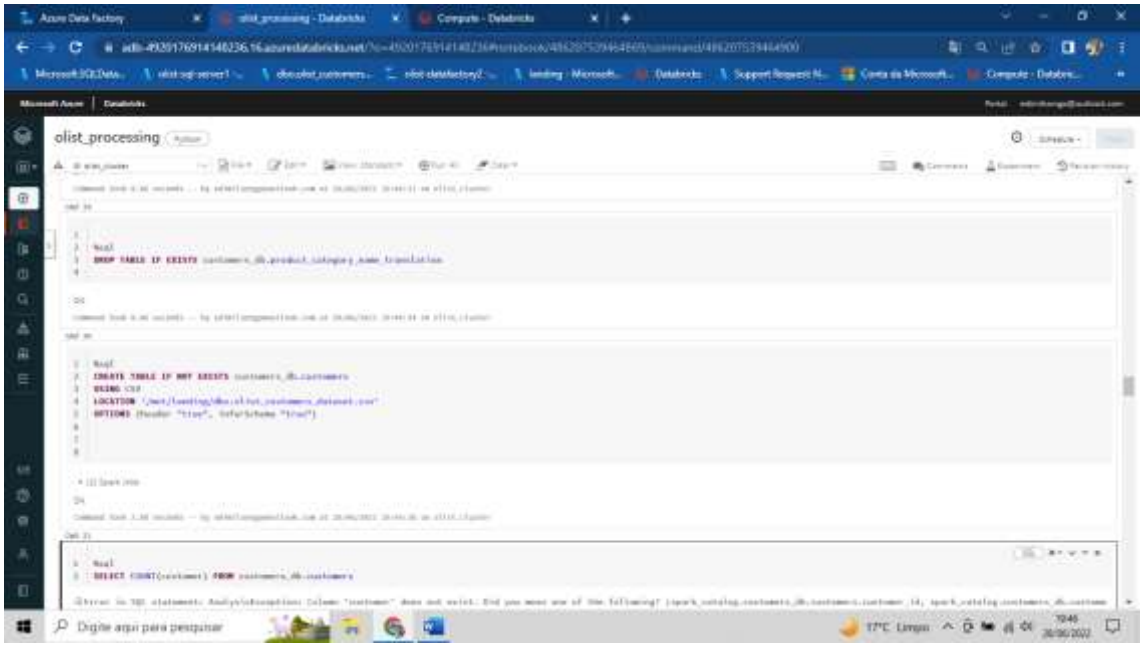
LOCATION '/mnt/landing/dbo.olist_customers_dataset.csv'

OPTIONS (header "true", inferSchema "true")


olist_processing
Create SQL Database
Create SQL Tables

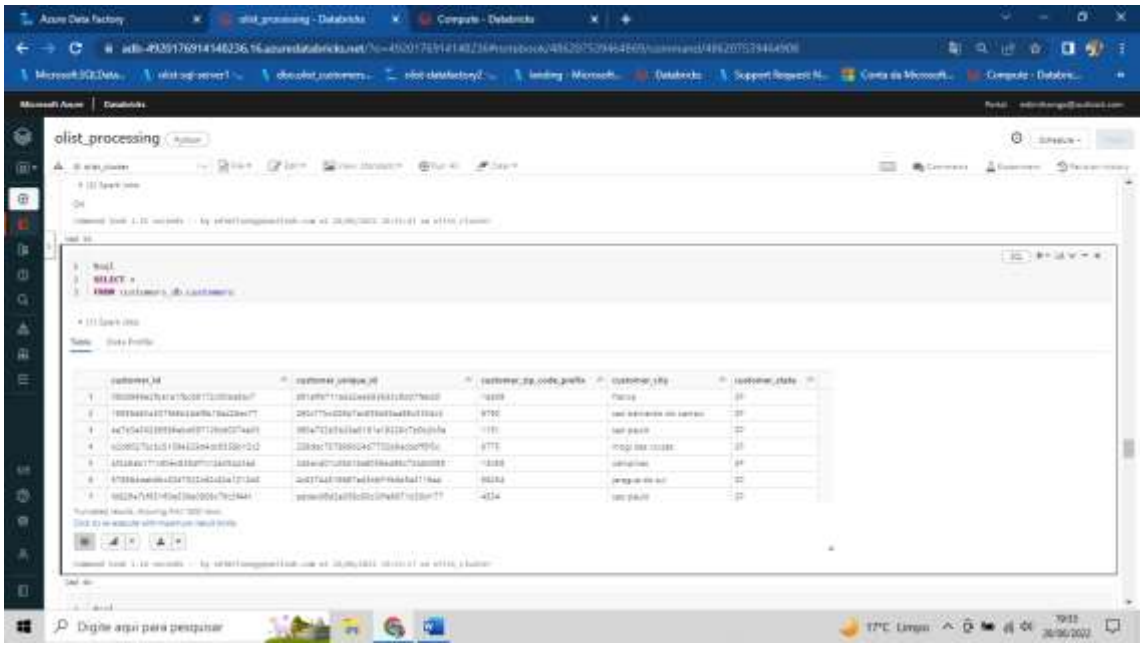

olist_processing

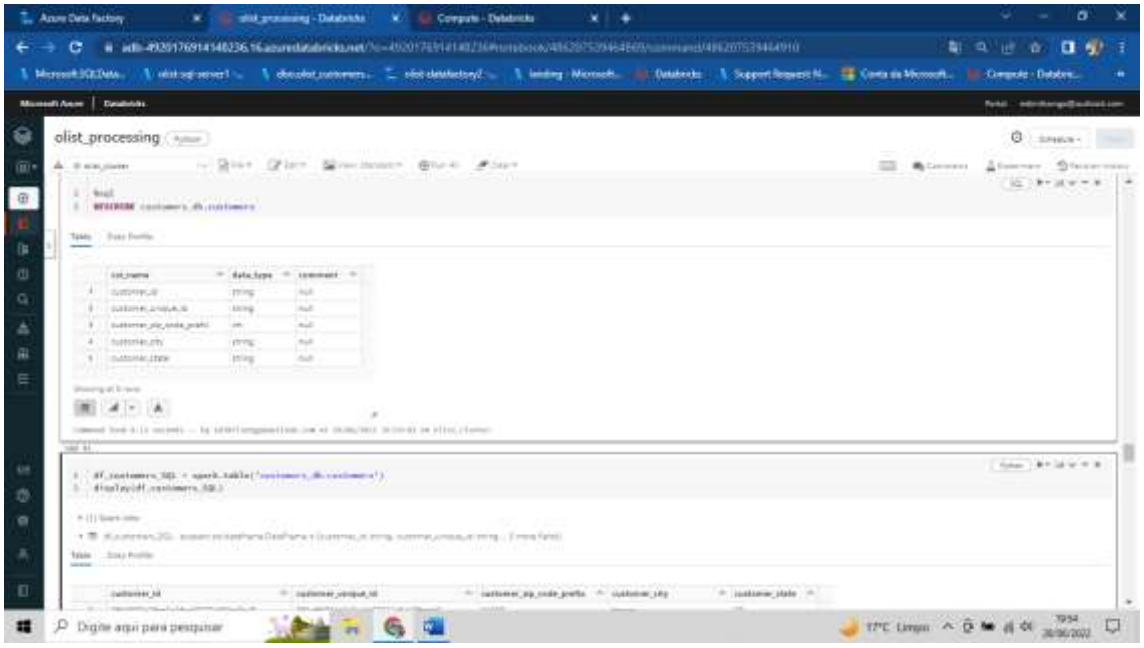




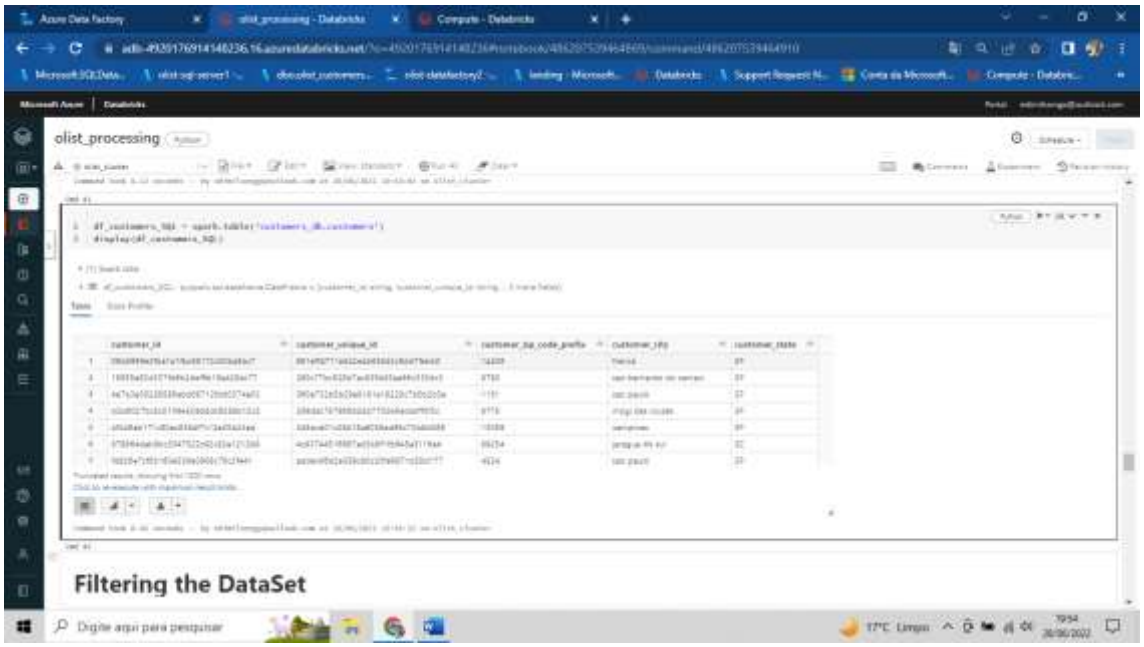
FAZER O MESMO COMANDO PARA AS DEMAIS TABELAS

AGORA VAMOS ACESSAR A TABELA

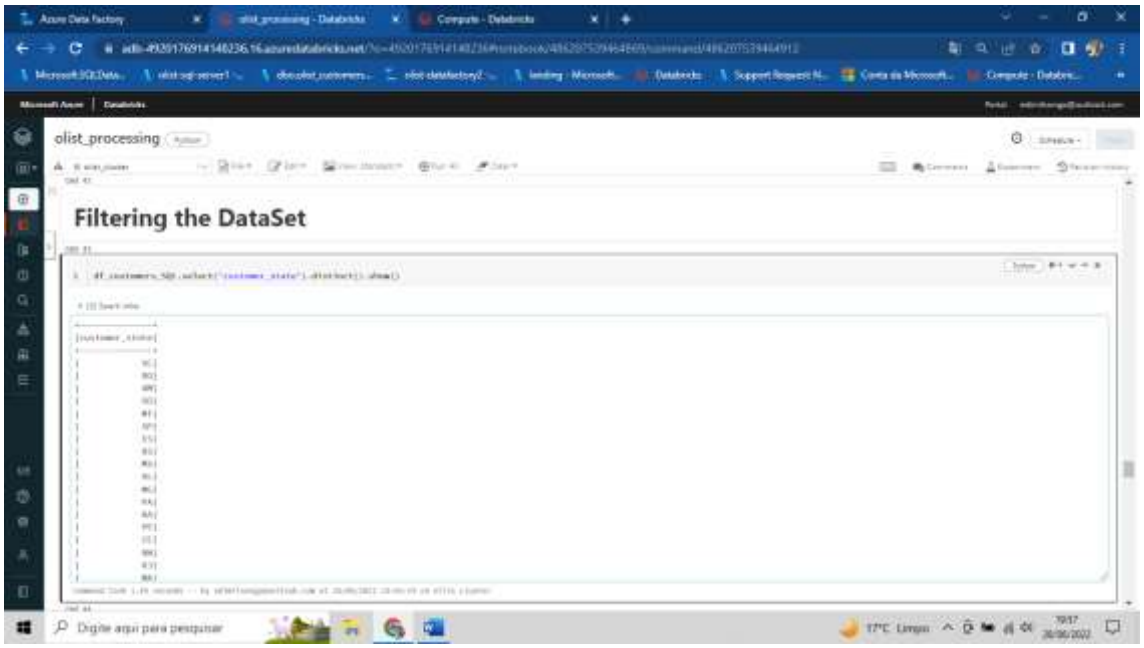




AQUI POSSO TRANSFORMAR O DATAFRAME EM SQL



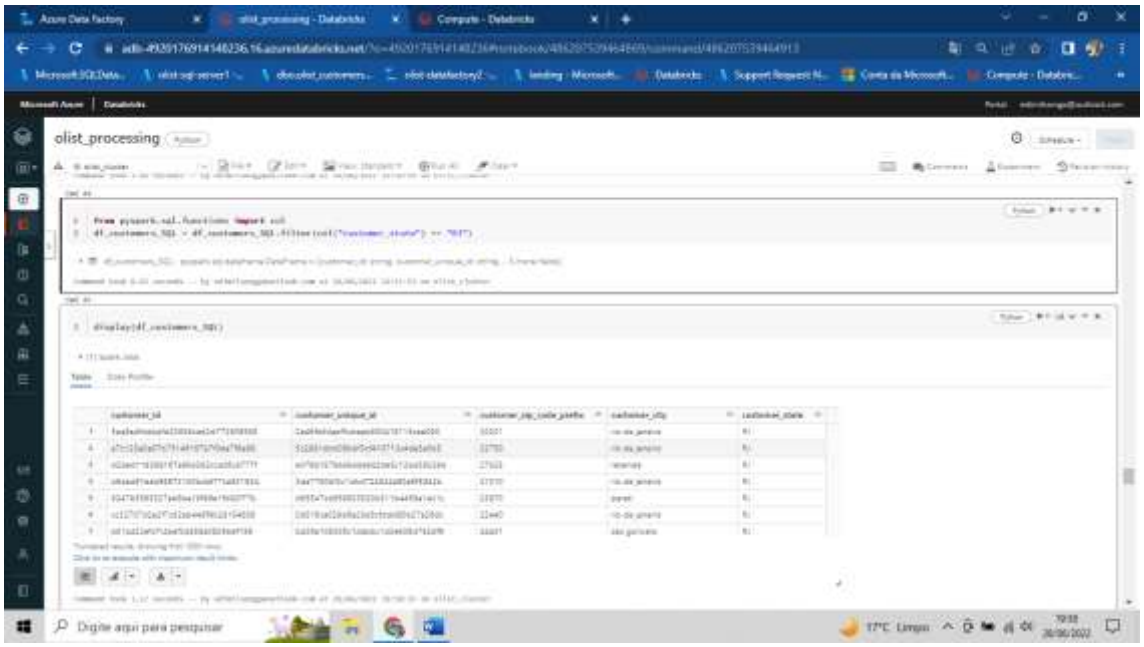
FILTRAR O DATAFRAME



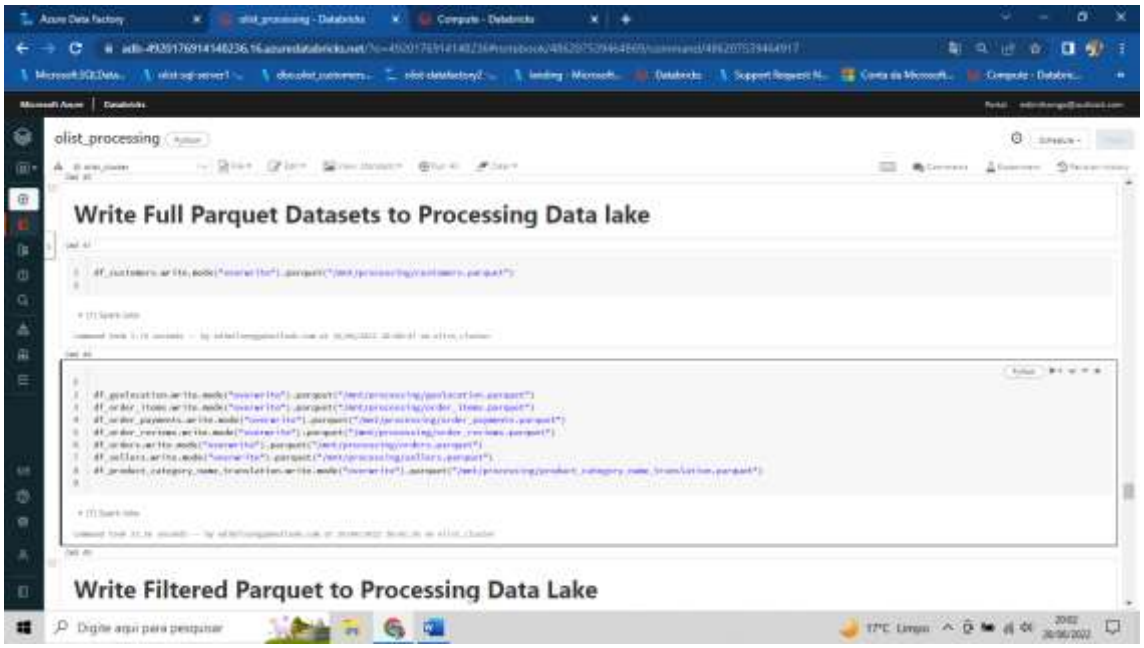
FILTRANDO O DATAFRAME PELA COLUNA customer_state == RJ

```
from pyspark.sql.functions import col

df_customers_SQL = df_customers_SQL.filter(col("customer_state") == "RJ")
```

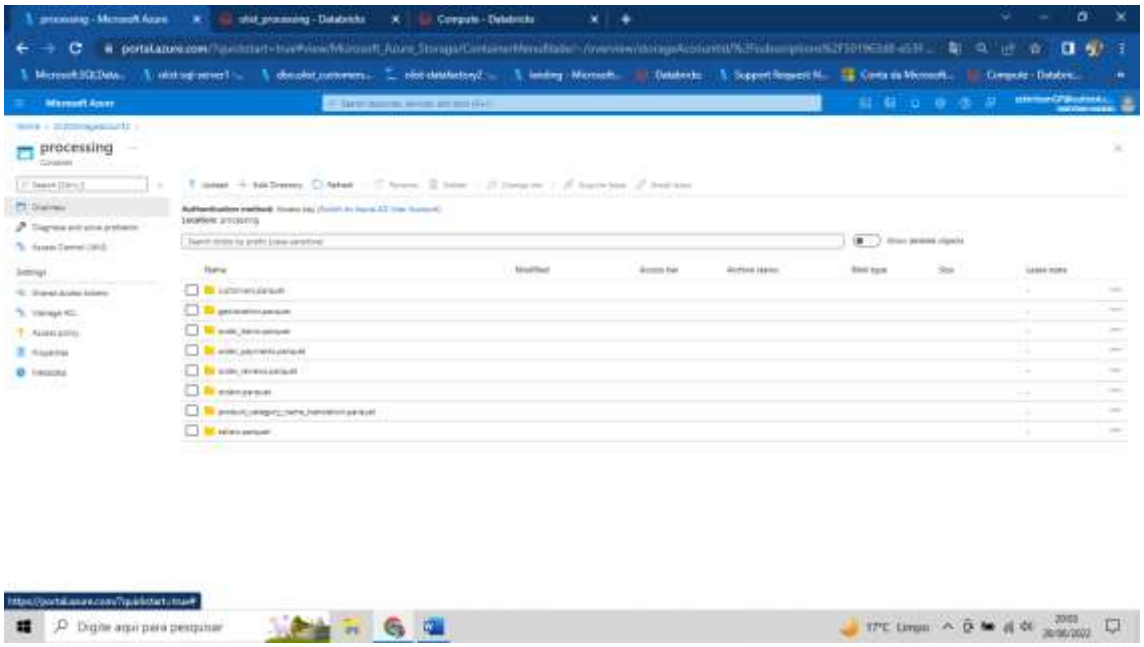


AGORA VAMOS LER E GRAVAR EM FORMATO PARQUET TODAS AS TABELAS



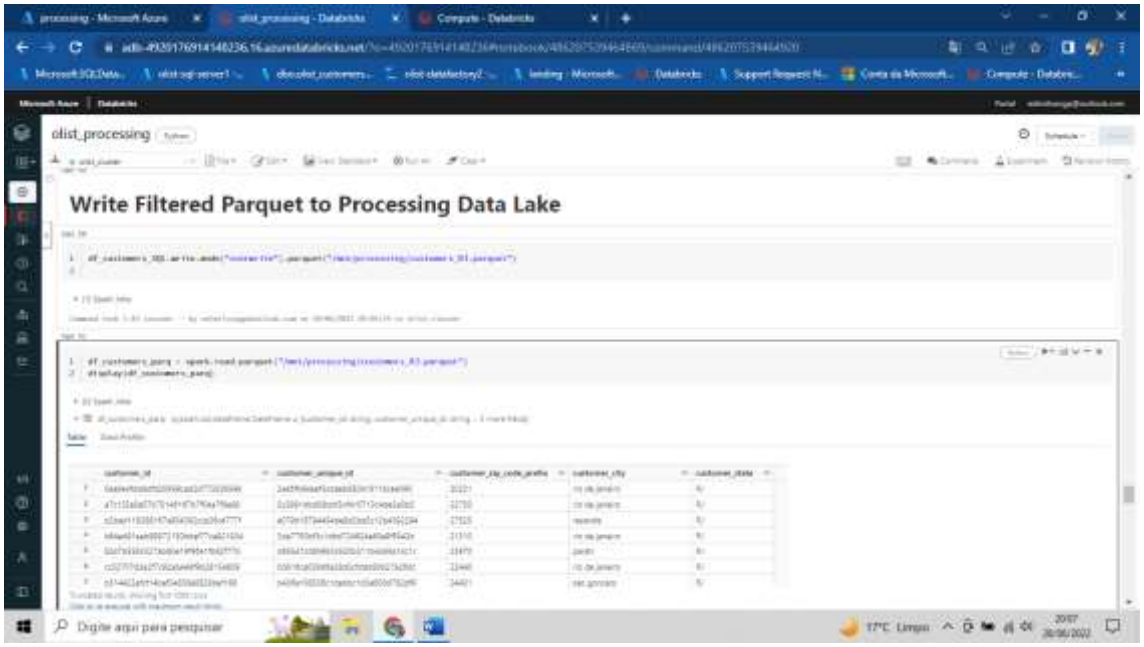
PARA VERIFICAR SE CARREGOU AS TABELAS EM STORAGE ACOUNT PROCESSING

TABELA CRIADAS COM SUCESSO EM FORMATO PARQUET



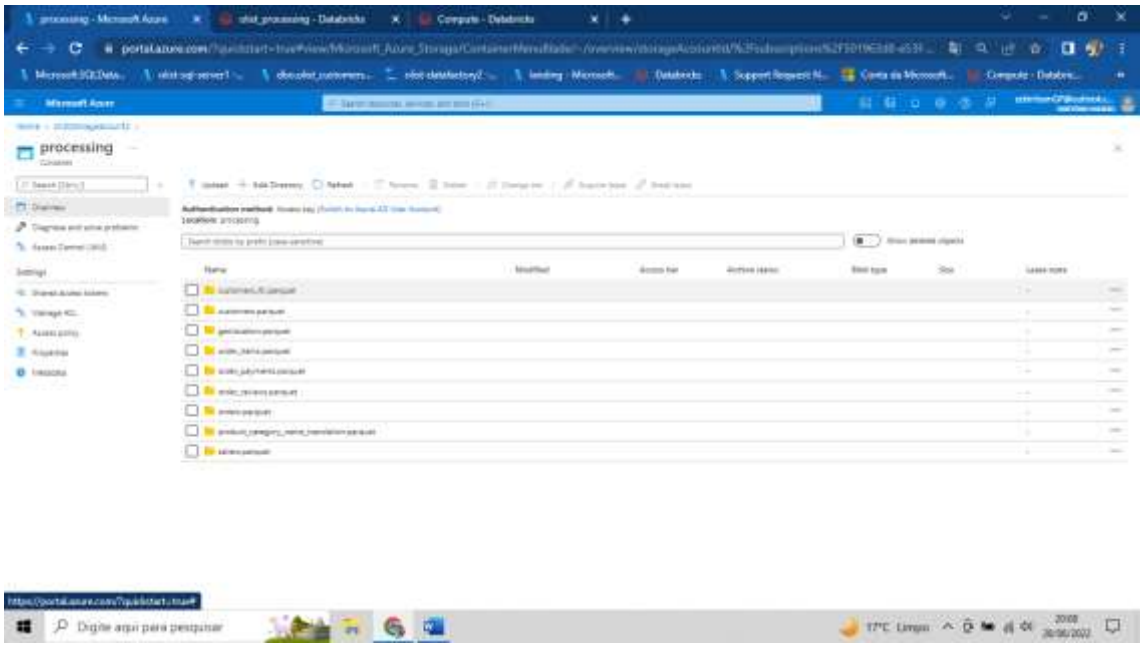
CASO EU QUEIRA SALVAR UM DATAFRAME JÁ FILTRADO

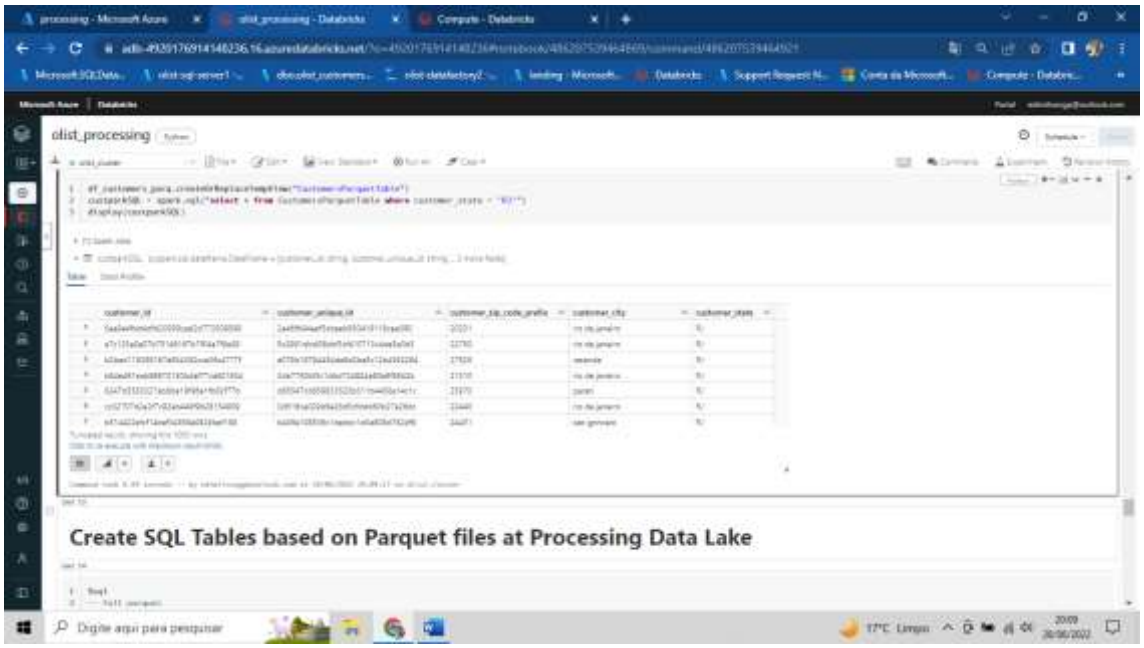
E VERIFICAR SE ESTÁ FUNCIONANDO CORRETAMENTE



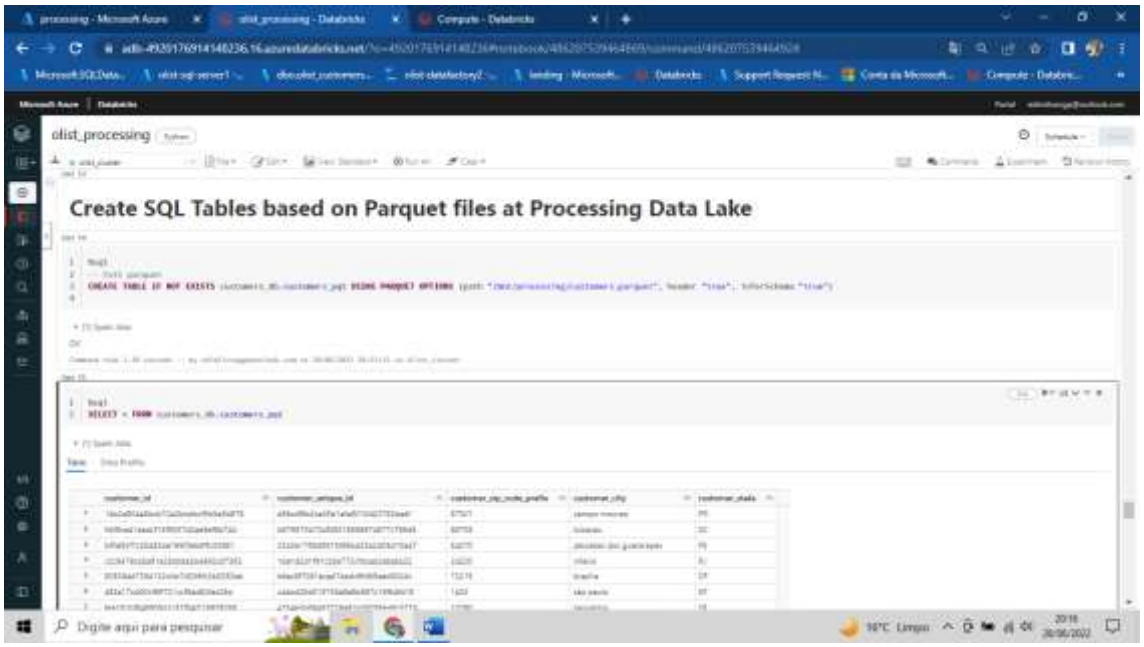
VERIFICANDO SE SALVOU EM PROCESSING

SALVOU COMO RJ.PARQUET O DATAFRAME FILTRADO

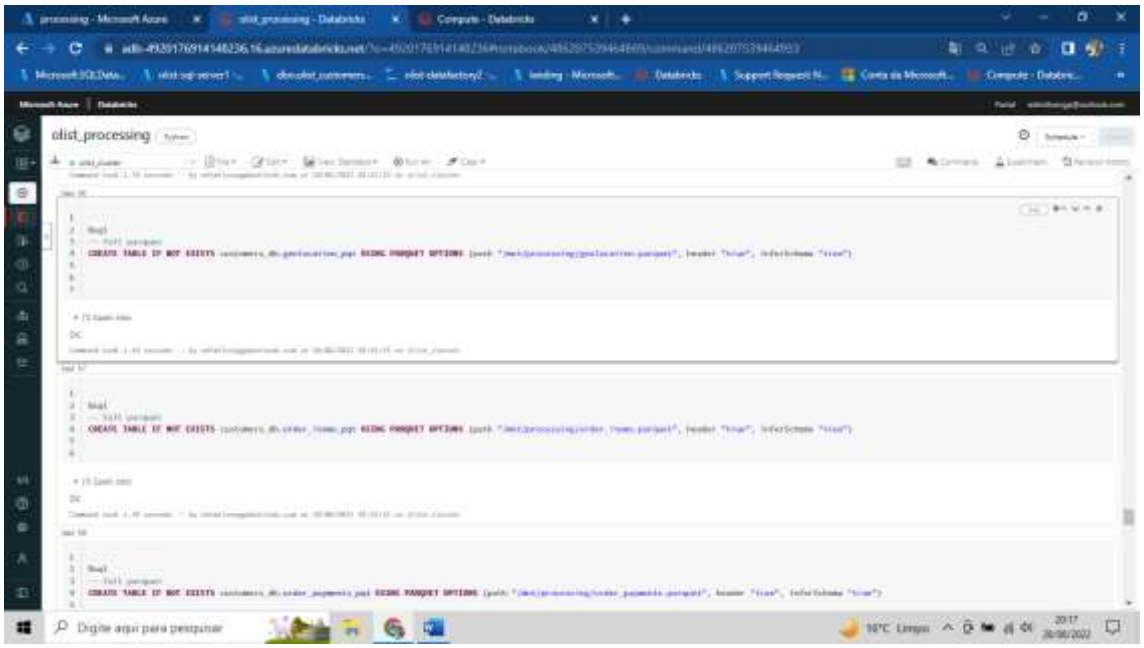




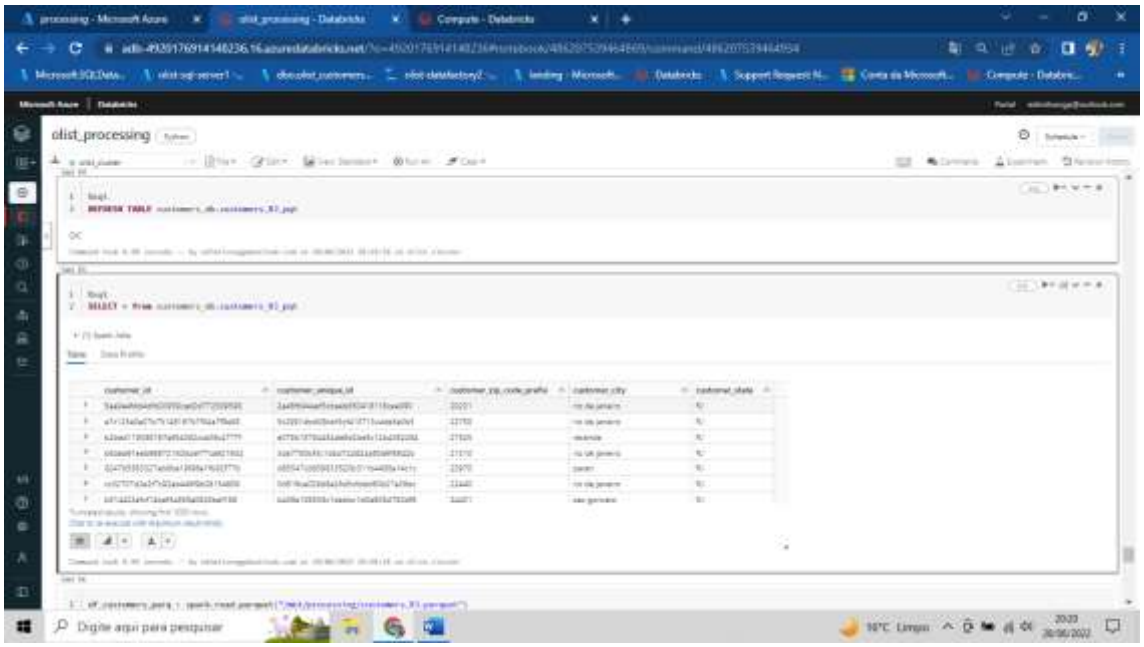
AGORA VAMOS CRIAR TABELAS BASEADAS NO PARQUET

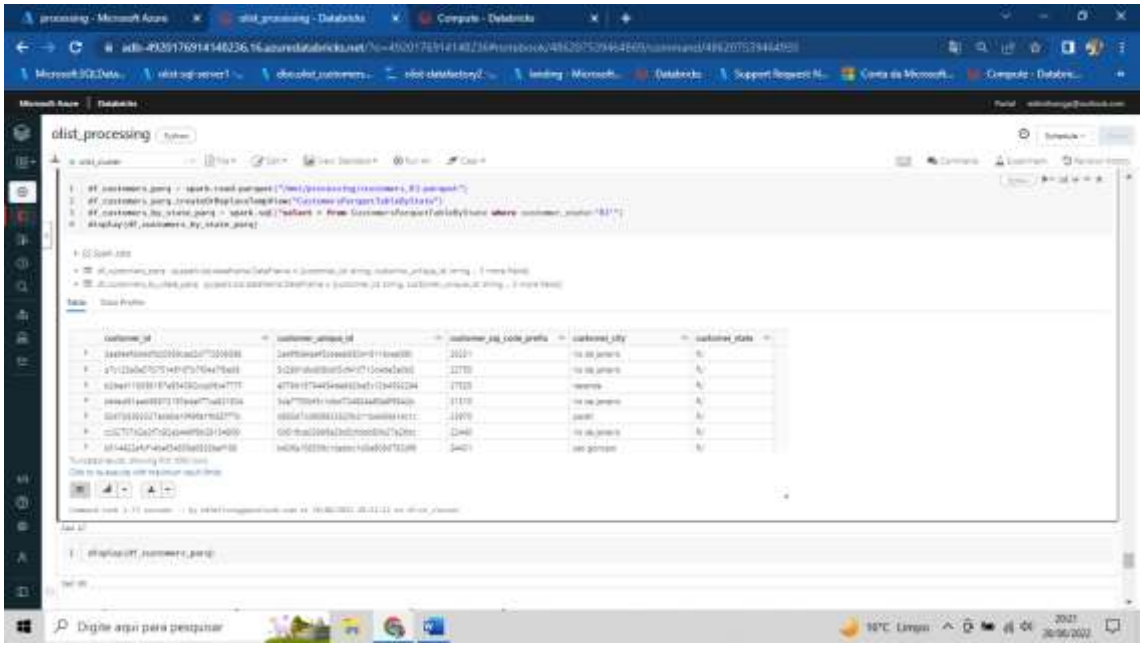


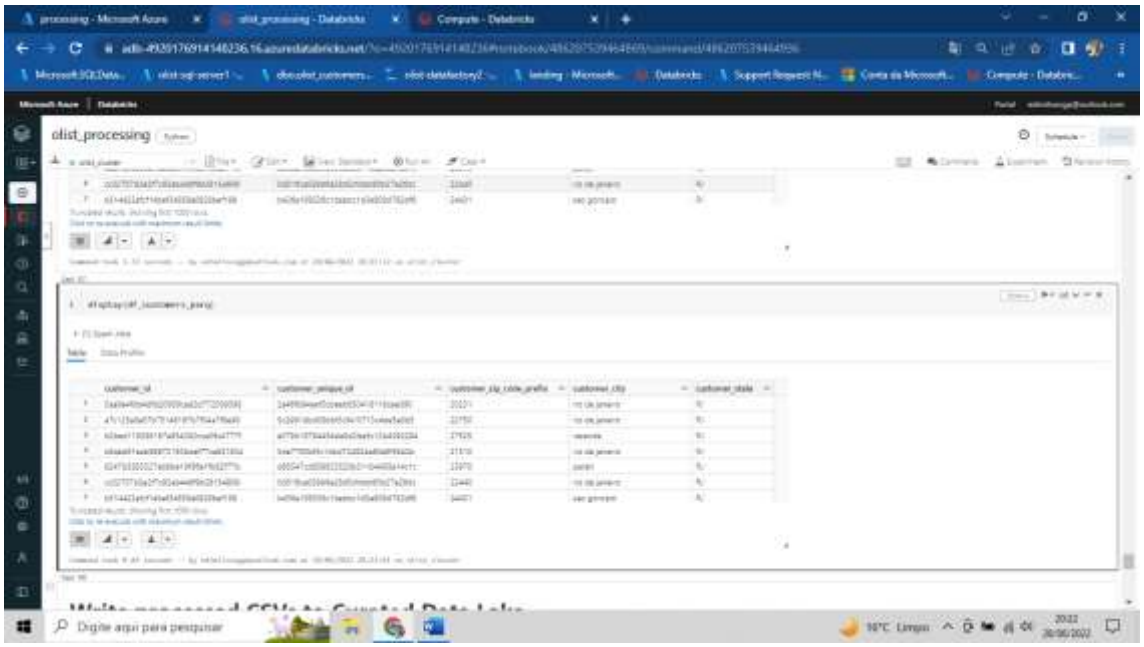
FAZER O MESMO PROCESSO PARA AS OUTRAS TABELAS



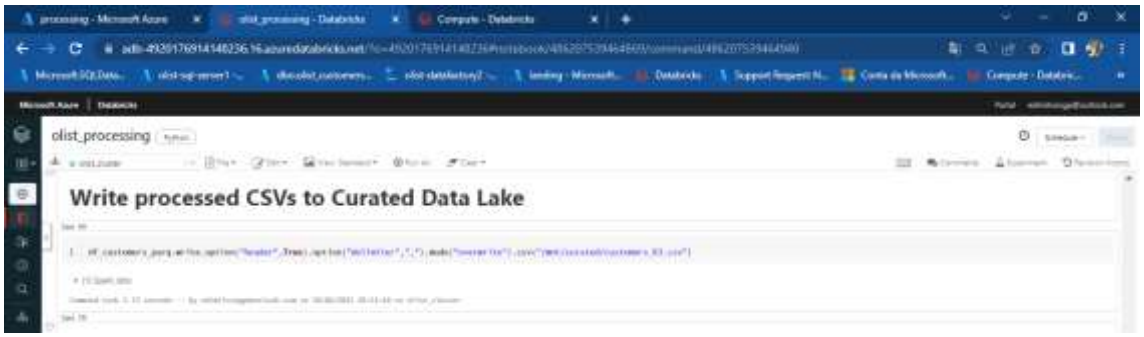
COMANDO REFRESH VAI BUSCAR OS DADOS MAIS RECENTES


olist_processing
Digite aqui para pesquisar

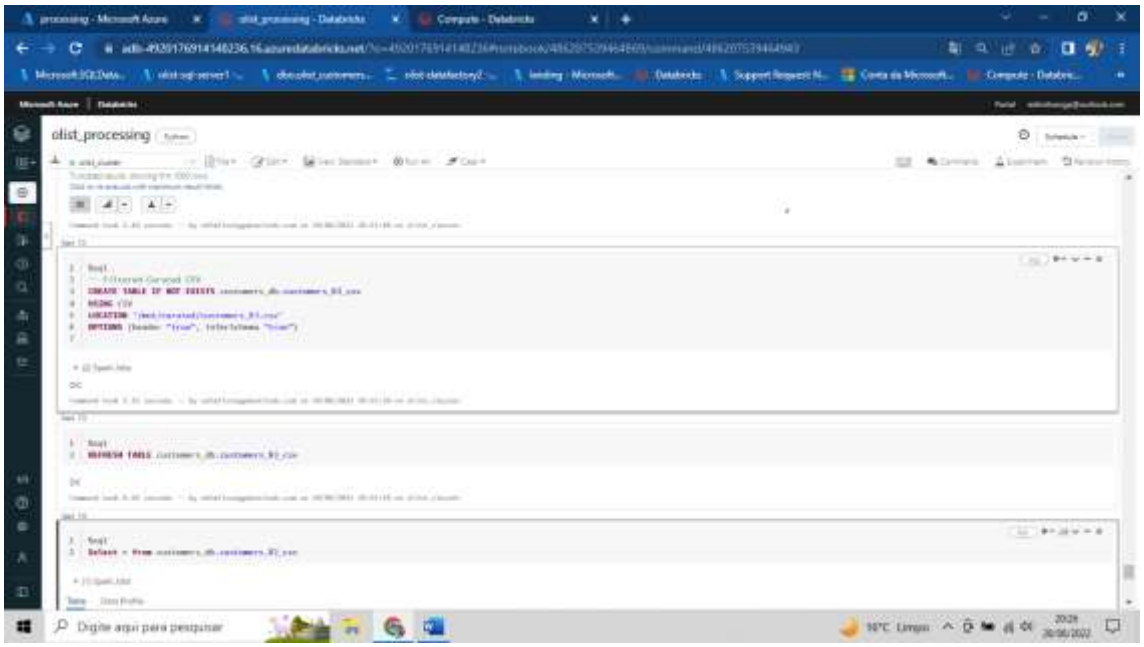

olist_processing
Digite aqui para pesquisar

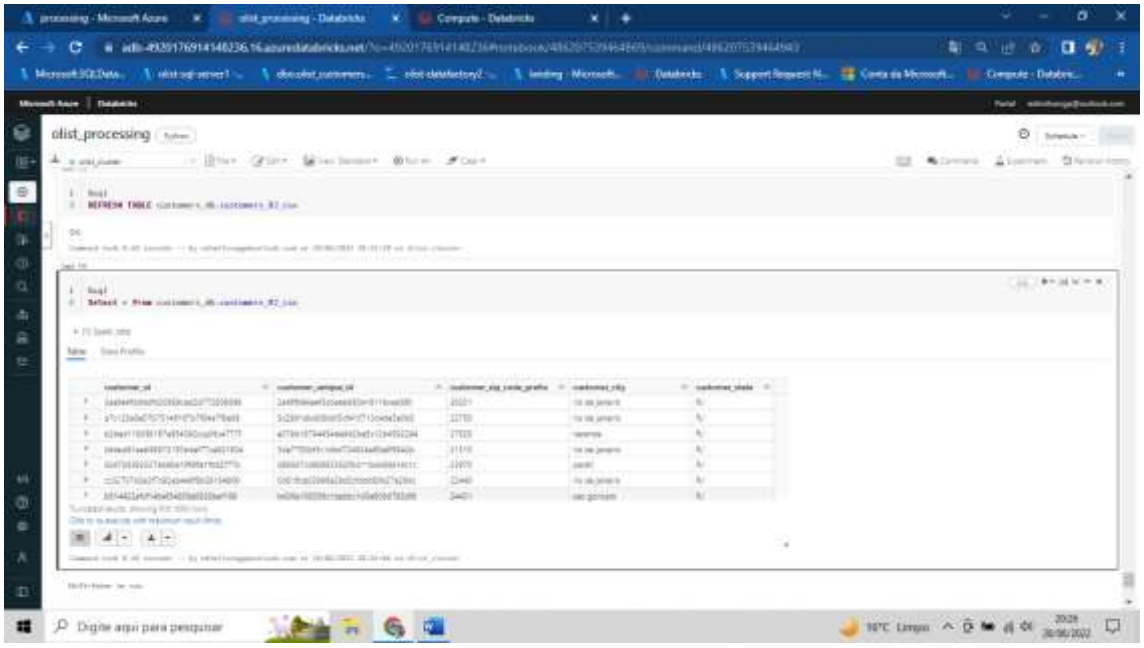


CRIANDO CSV NA CURATED



VAMOS FAZER O TESTE E LER O ARQUIVO

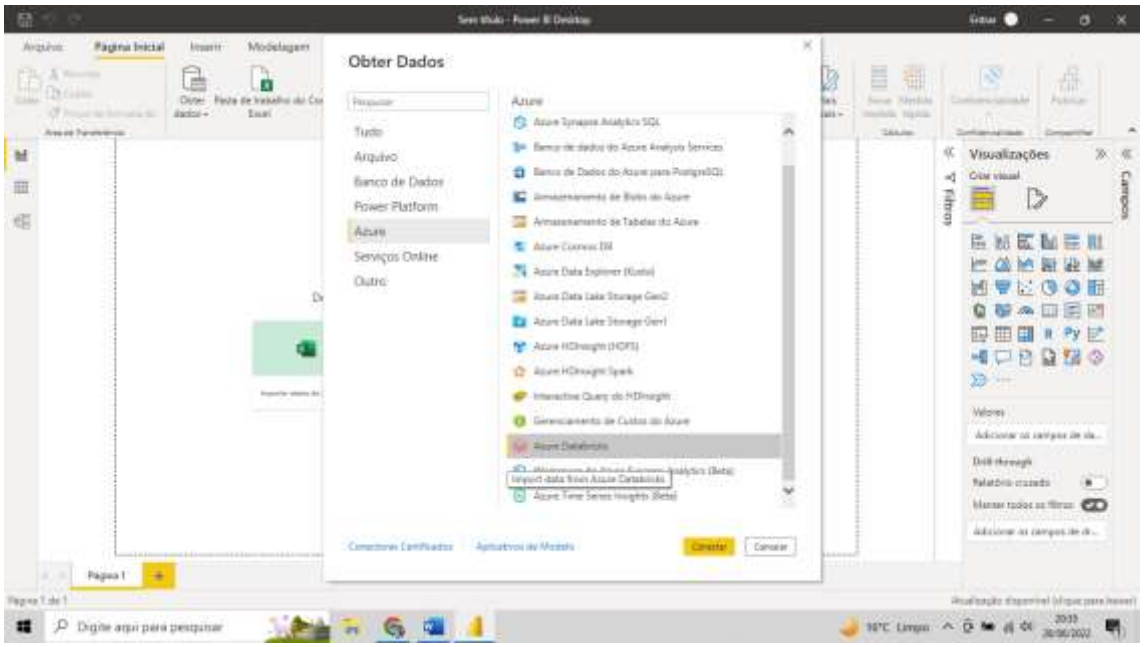




VERIFICANDO A CRIAÇÃO DO CSV FILTRADO EM CURATED

ARQUITETURA ESTÁ PRONTA AGORA LEVAR PARA O POWER BI

CONECTANDO DO BANCO DE DADOS DO AZURE DATABRICKS

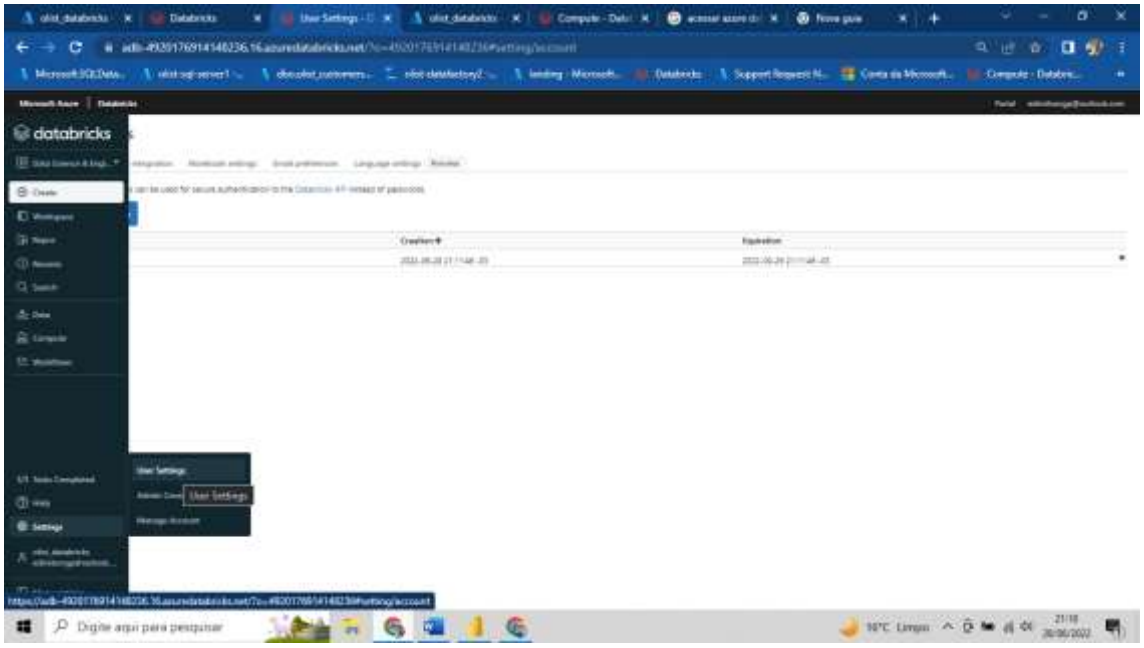


PARA PEGAR ESSAS INFORMAÇÕES ABAIXO

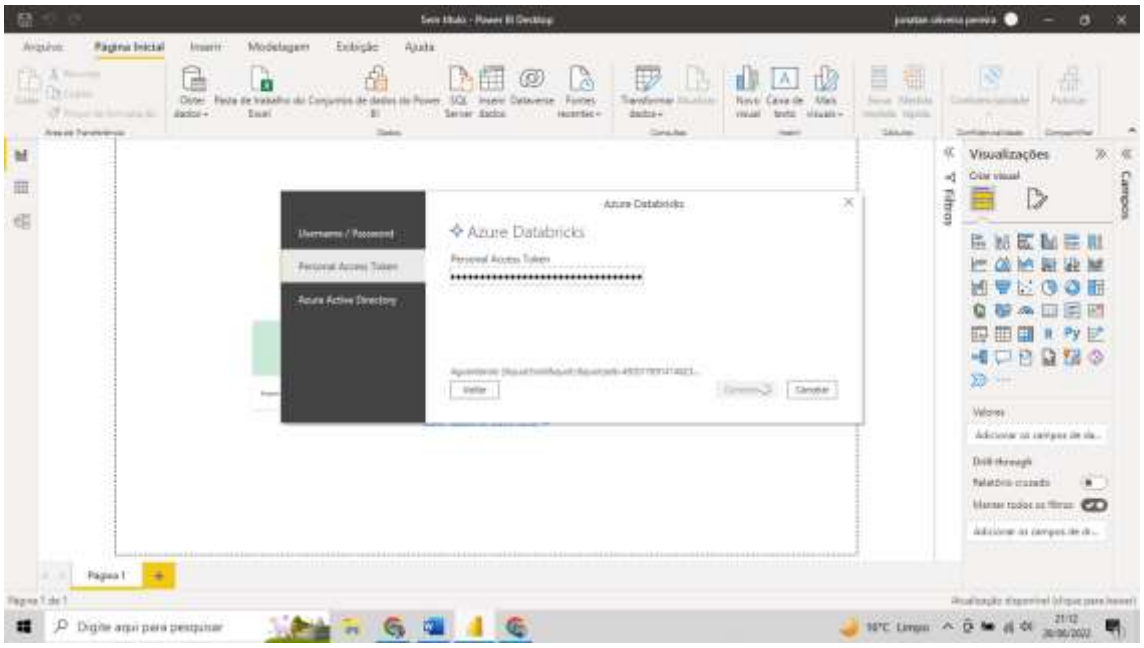
DIRECT QUERY -> PARA PEGAR AS INFORMAÇÕES DO DATA LAKE EM TEMPO REAL

AS INFORMAÇÕES ESTÃO NO CLUSTER DO DATABRICKS

Para gerar o token

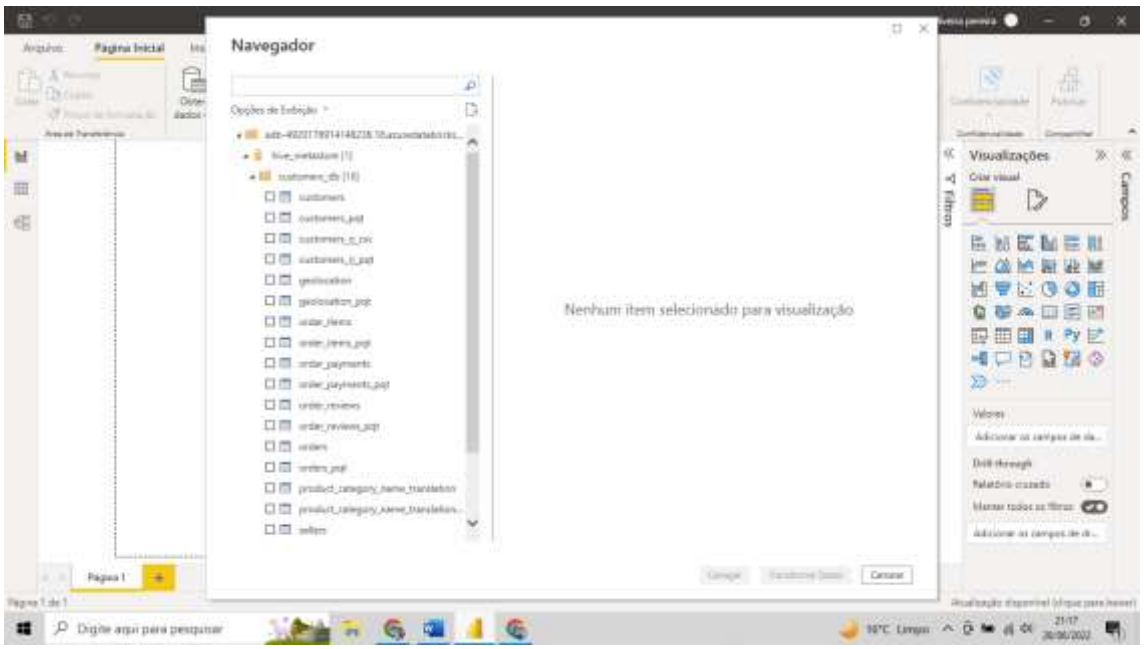


Conectando por token

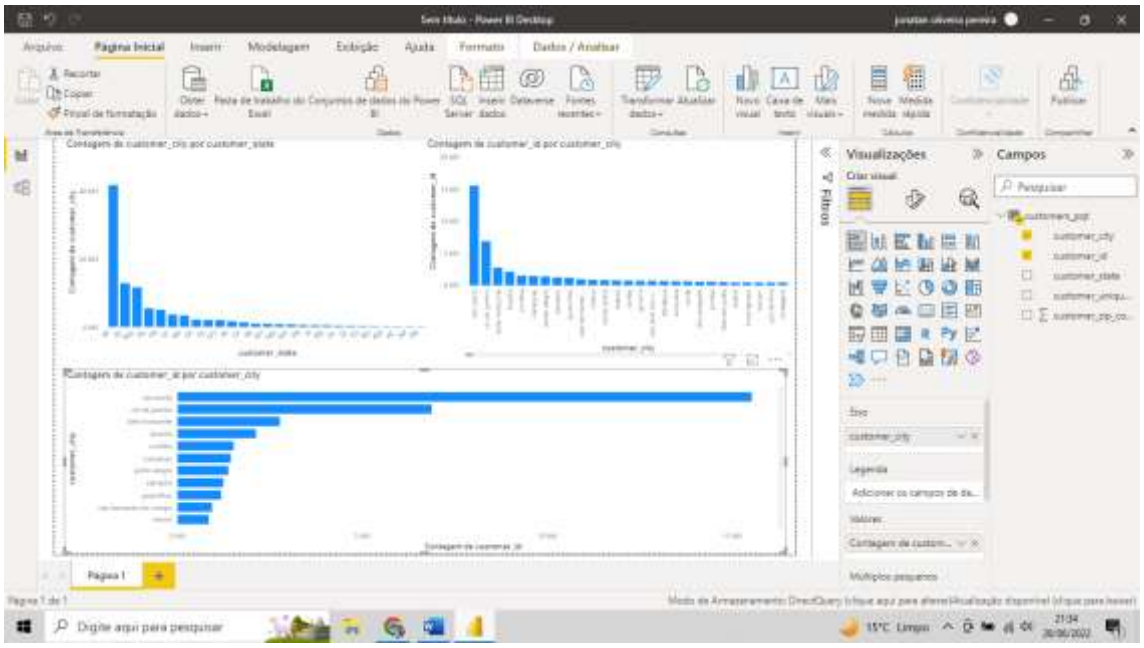


Acesso pelo token funcionando

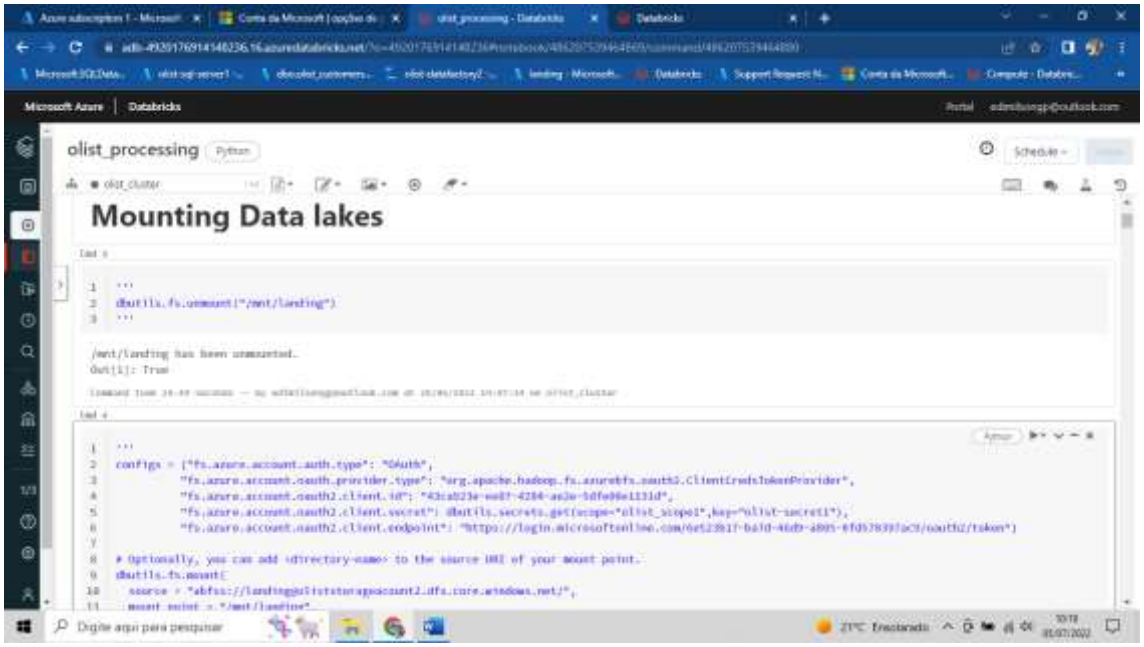
POWER BI – aqui deve ser feito as Analises



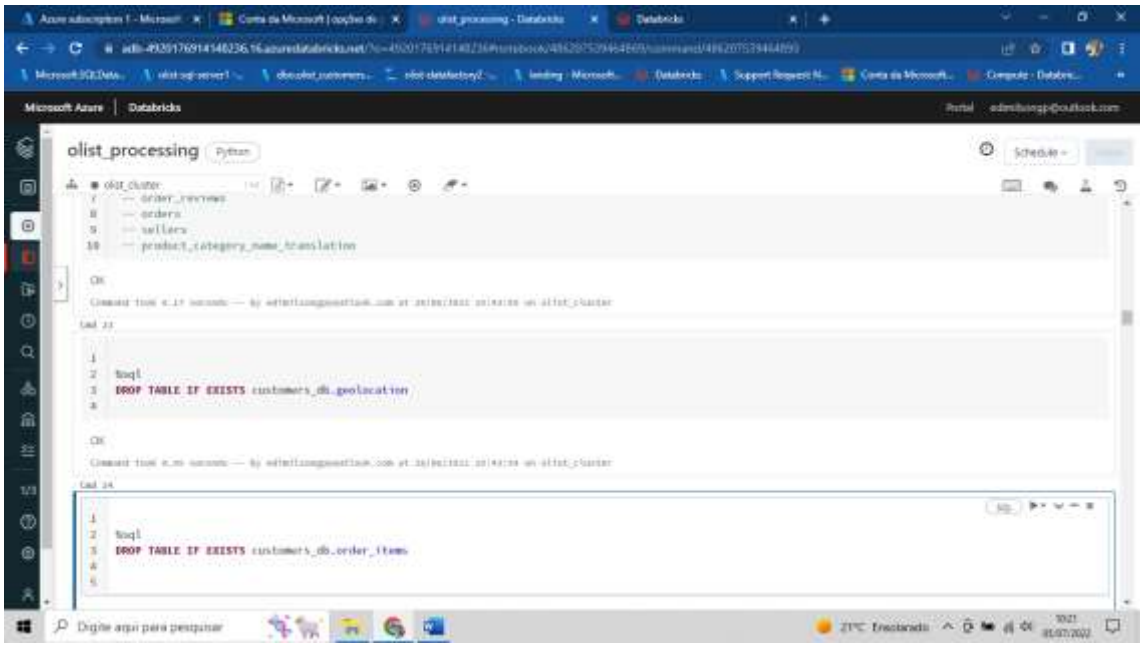
TESTE COM POWER BI



FINALIZANDO A ARQUITETURA DE DADOS – PREPARANDO O NOTEBOOK PARA EXECUÇÃO AUTOMATICA COMENTAR TODAS AS LINHAS DE MOUNTAGEM (mnt)



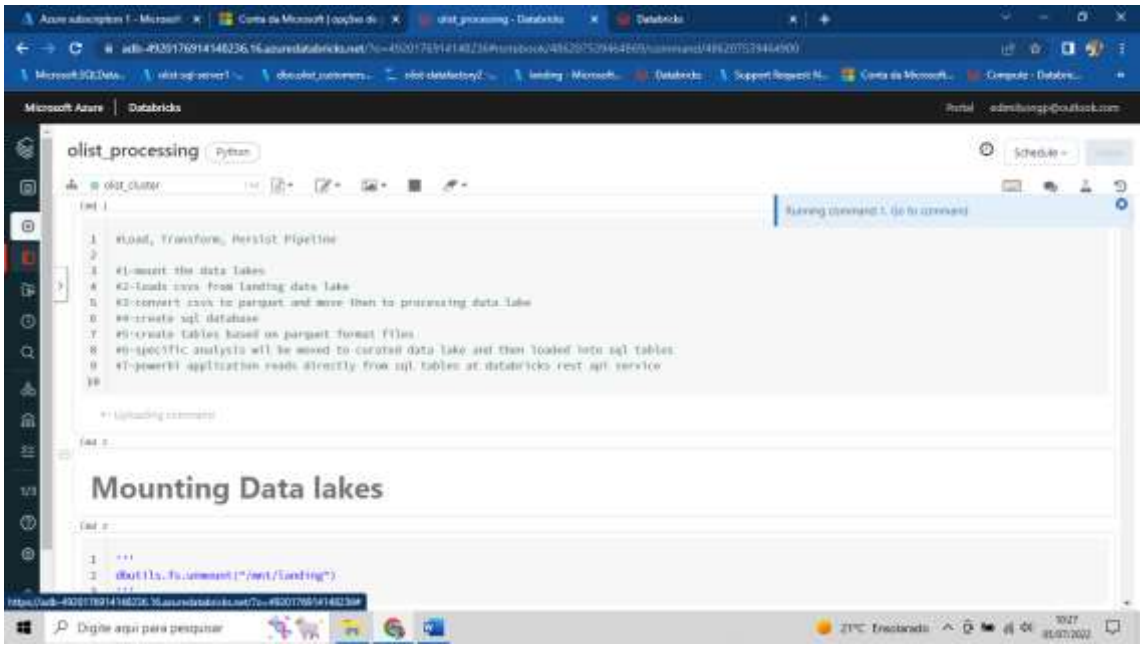
Descomentar todos os DROP e as demais a baixo



Agora vamos ativar o Cluster

Para executar novamente o Script

Executando o Script

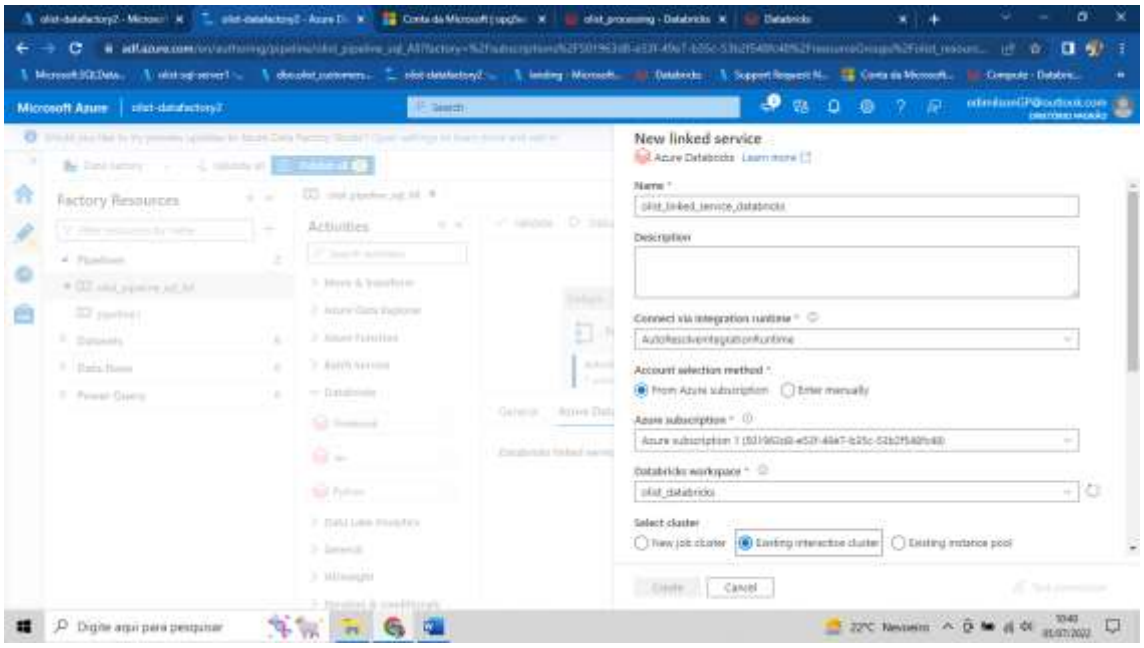


FAZER O AGENDAMENTO DO NOTEBOOK(SCRIPT) PARA RODAR AUTOMATICAMENTE DOM DATA FACTORY

EM DATABRICKS TEM O NOTEBOOK

CLICAR E ARRASTAR PARA AREA

CONFIGURANDO

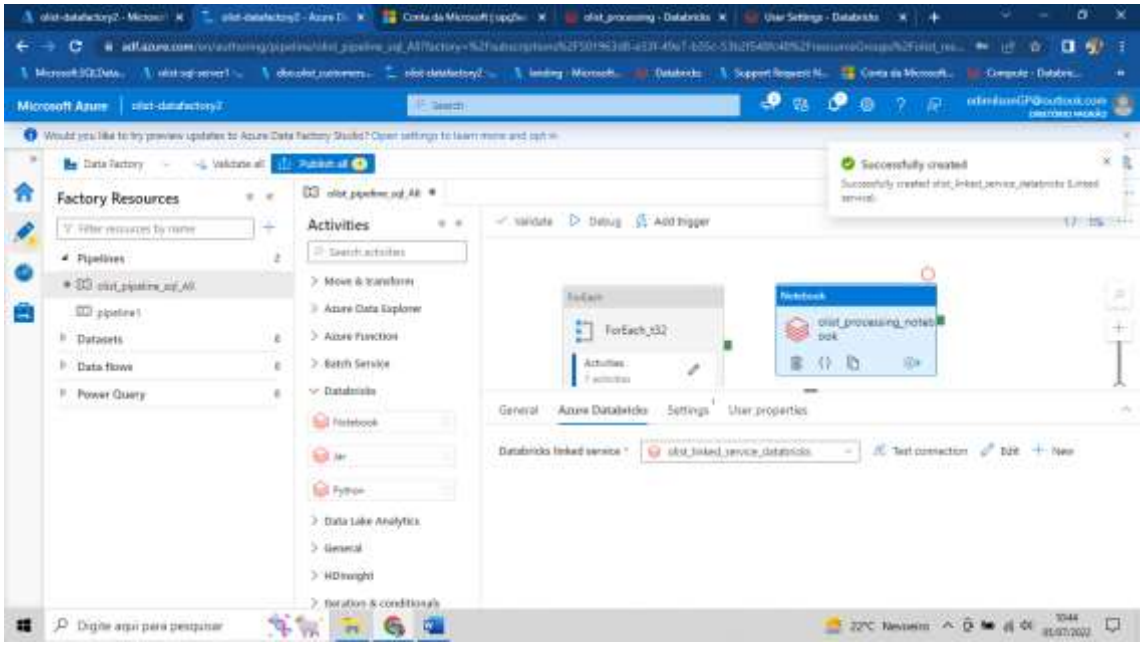


CRIAR TOKEN DE ACESSO

AGORA COPIAR

CLICAR EM CRIAR

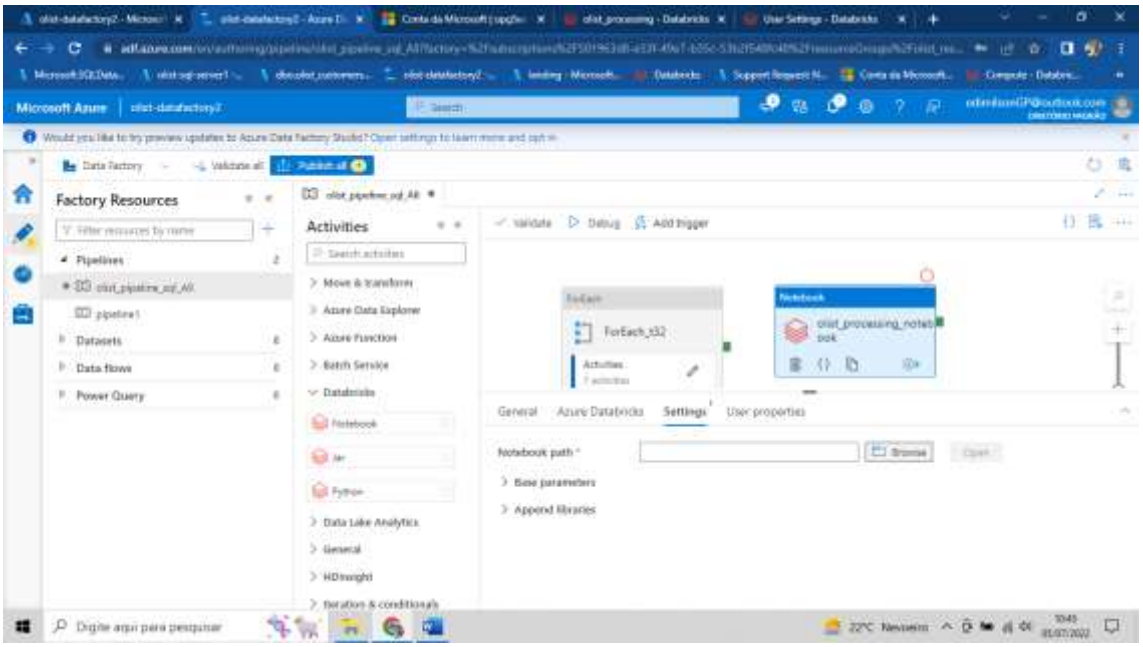
CRIADO



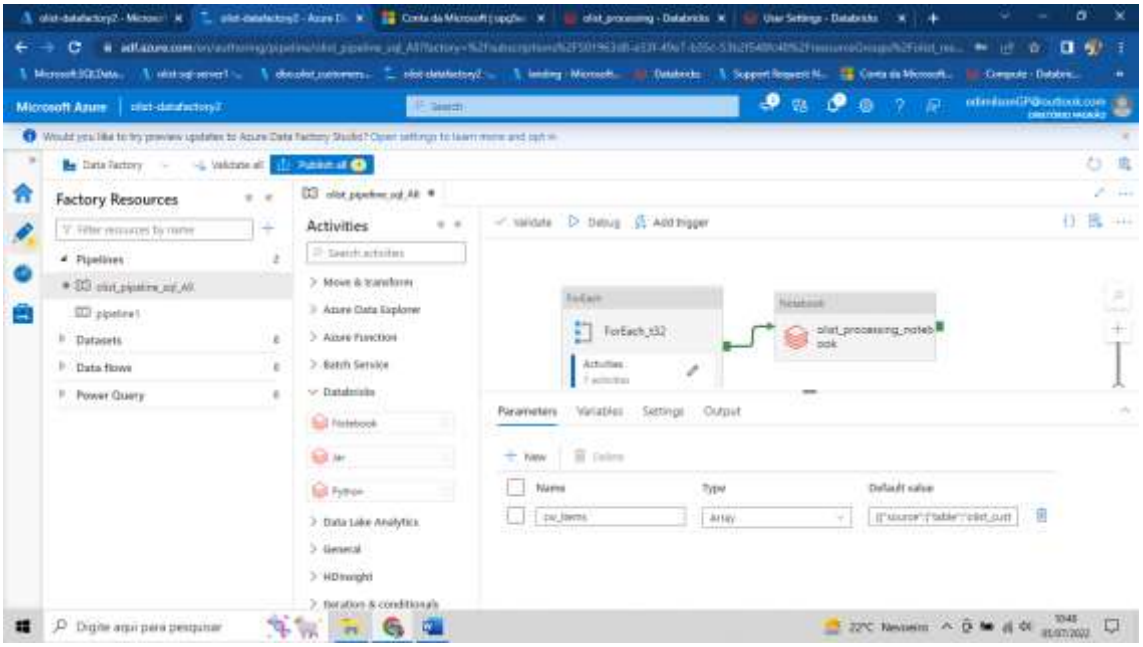
TESTE DE CONEXÃO

AGORA CONECTAR O NOTEBOOK

EM CONFIGURAÇÃO



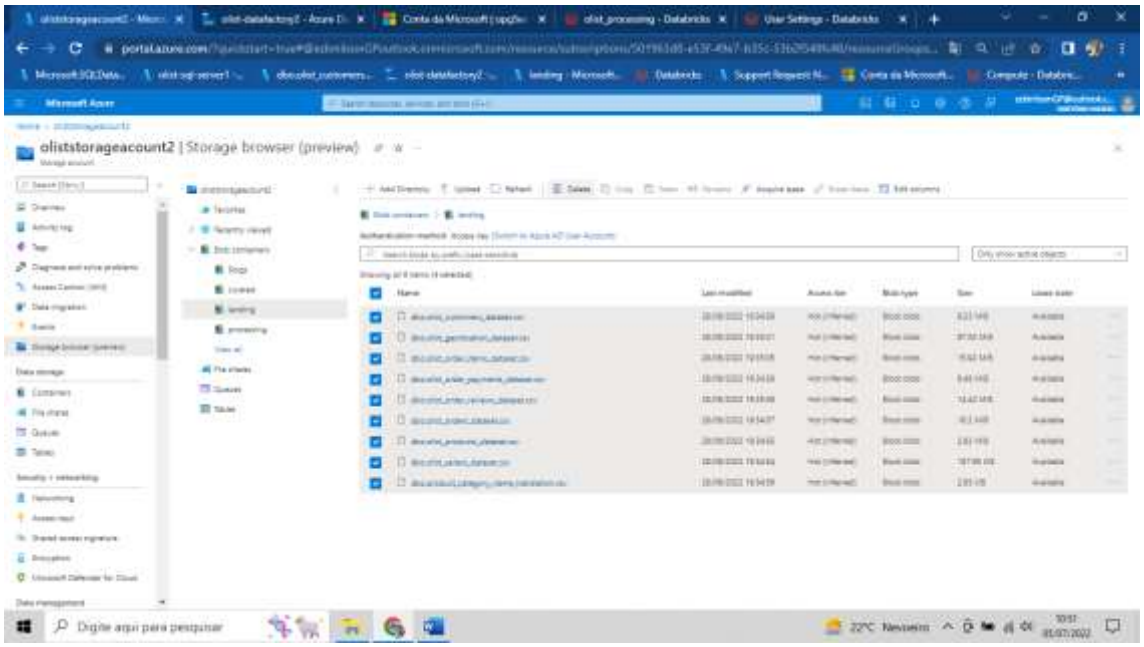
CONECTANDO



AGORA VAMOS FAZER O TESTE

VAMOS DELETAR

EM LANDING

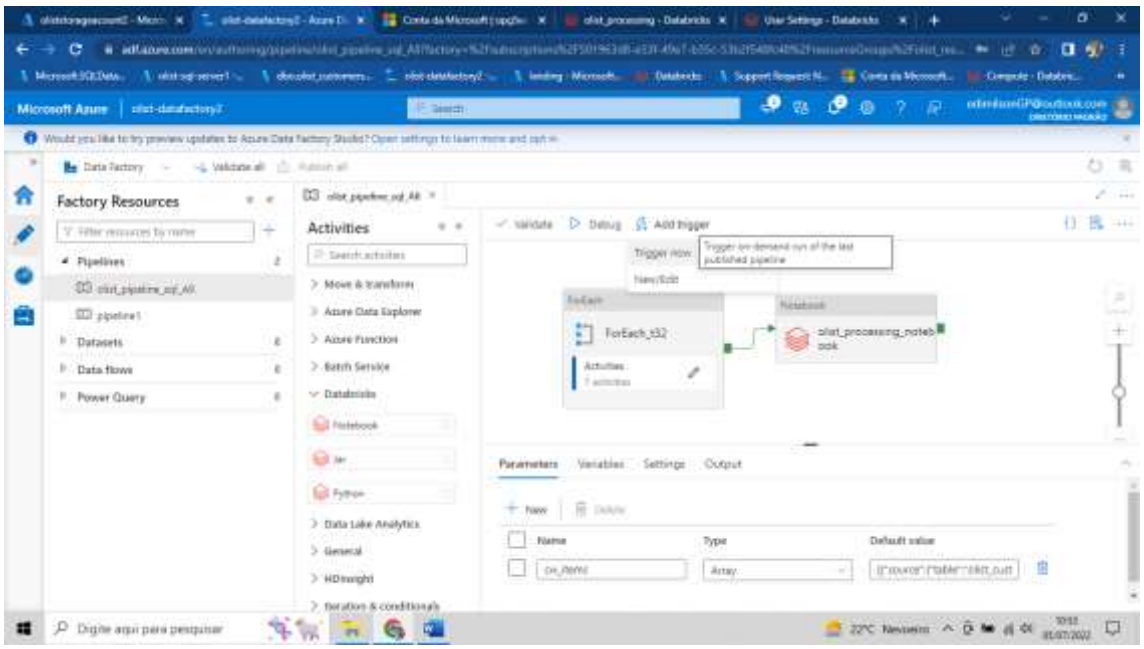


DELETADO

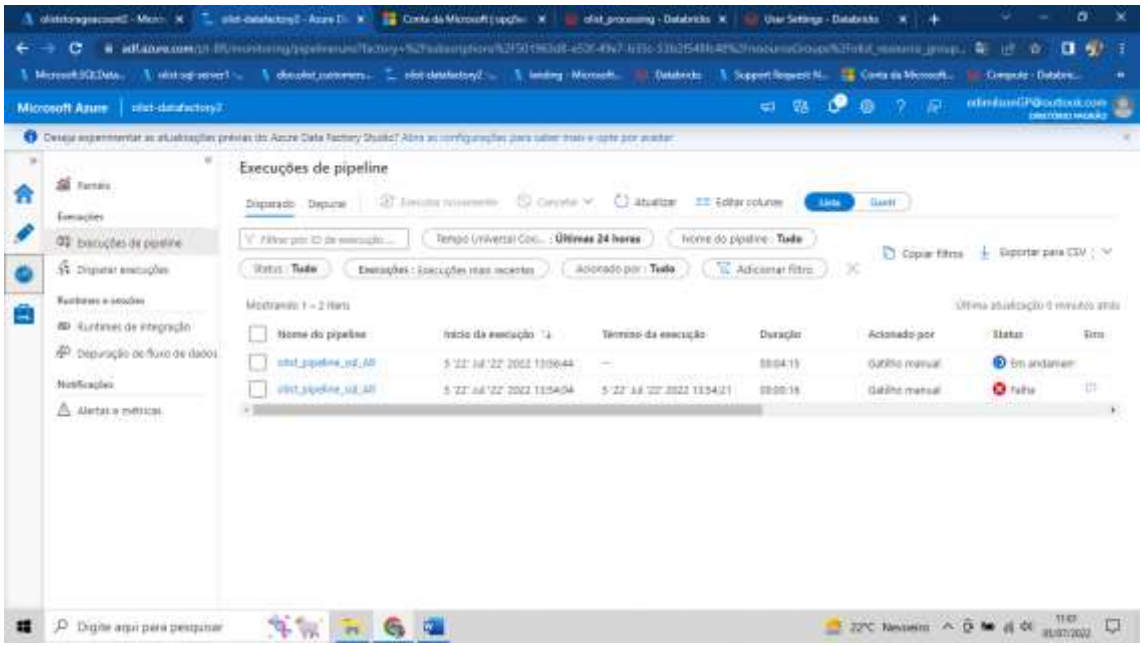
FAZER ISSO PARA OS DEMAIS CONTEINERS

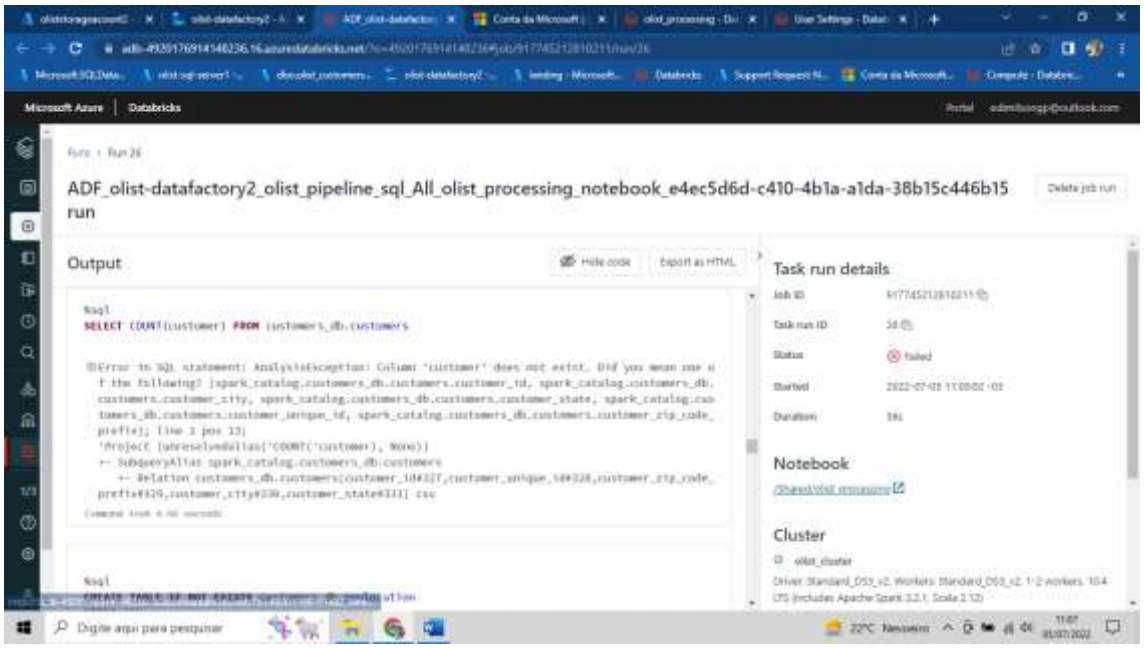
PROCESSING E CURATED

AGORA NO DATA FACTORY VAMOS DISPARAR O GATILHO



APRESENTOU ERRO



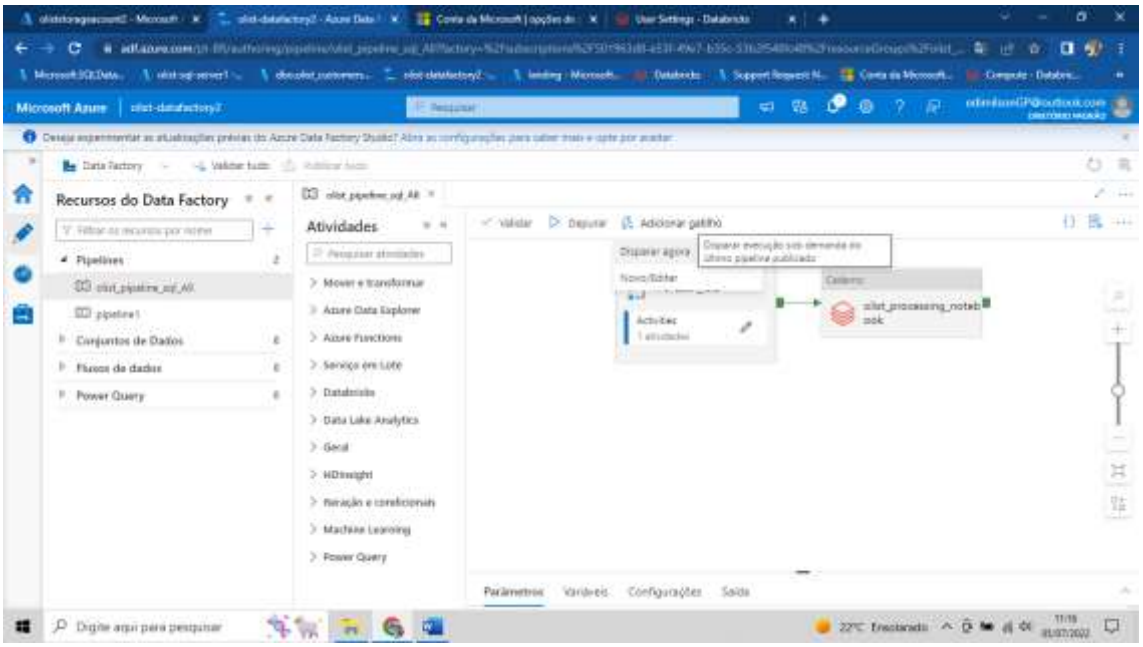


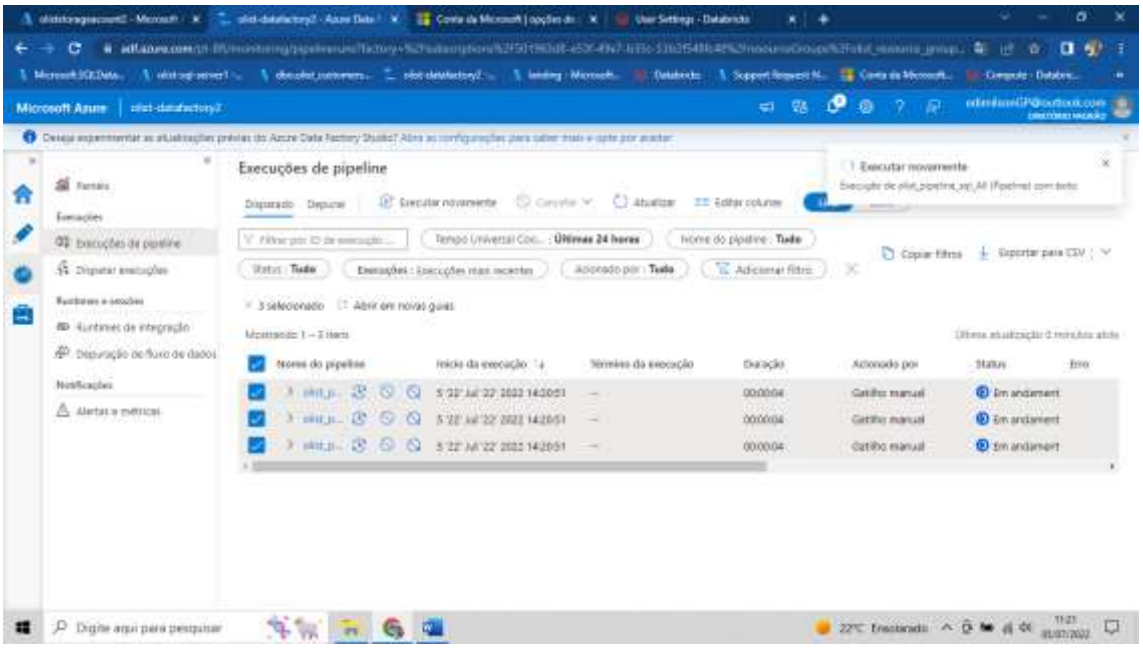
CORREÇÃO REMOVI O COUNT

VERIFICAR SE TEM O DBO NA FRENTE TAMBÉM DE TODOS OS LOCATION

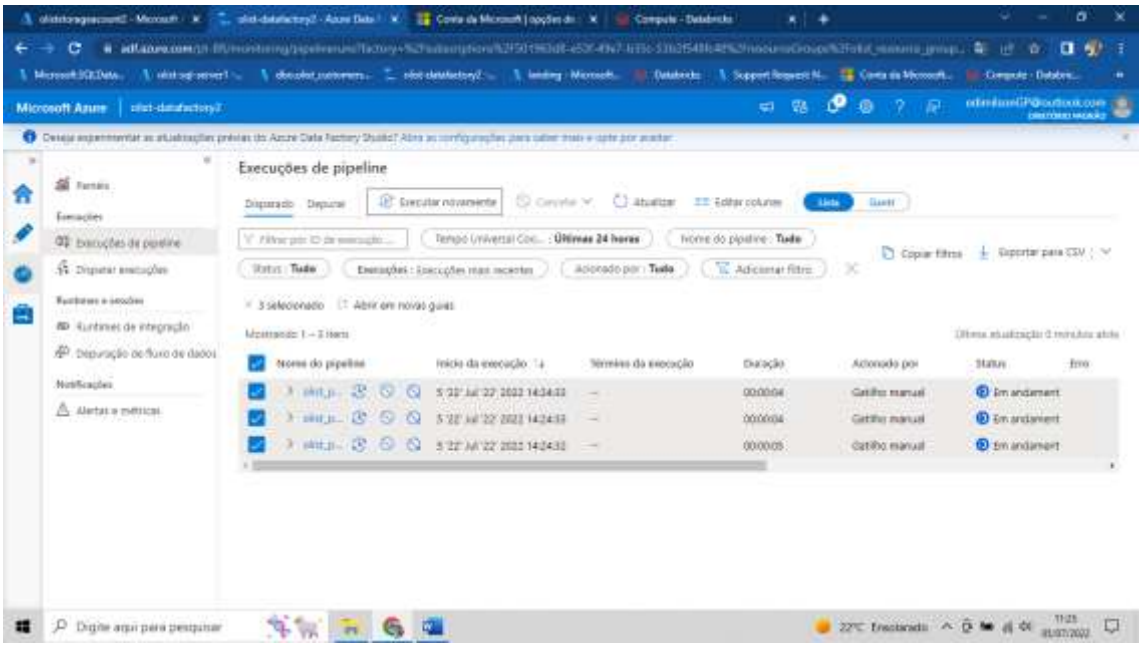
E FECHAR O NOTEBOOK

EXECUTAR O PIPELINE NOVAMENTE NO DATA FACTORY





EM ANDAMENTO


Execuções de pipeline
Disparado
Depurar
Executar novamente
Atualizar
Editar colunas
Lista
Gantt
Tempo Universal Coo... : Últimas 24 horas
Nome do pipeline : Tudo
Status : Tudo
Execuções : Execuções mais recentes
Acionado por : Tudo
Adicionar filtro
Copiar filtros
Exportar para CSV
3 selecionado
Abrir em novas guias
Nome do pipeline
Início da execução
Término da execução
Duração
Acionado por
Status
Erro
Gatilho manual
Em andamento
Execuções de pipeline
Execuções de pipeline
Disparar execuções
Runtimes de integração
Depuração de fluxo de dados
Alertas e métricas
Digite aqui para pesquisar