Crie seu pipeline de dados sem pagar por isso!

Erika Nagamine

23.03.21 12h00


TDC
ONLINE
Sala
ORACLE
Inovação
com dados
em nuvem
CRIE SEU PIPELINE DE DADOS
NA CLOUD SEM PAGAR
NADA POR ISSO!

Índice

1. Considerações iniciais e pré requisitos

A técnica demonstrada aqui no presente documento é apenas uma das que pode ser utilizada para criar um pipeline de dados. É possível utilizar outras técnicas dependendo do seu caso de uso. Para deixar o passo a passo de uma maneira única, padronizamos as telas em inglês.

Recursos usados:

OCI (all free tier)

- Armazenamento Oracle Object storage
- Functions (30 days free)
- Events (30 days free)
- Pacotes Oracle Instant Client + Tools e sqlplus
- Banco de dados autonomous (adw)
- Compute VM 2.1 Shape

Local

- Navegador

Tópicos não cobertos:

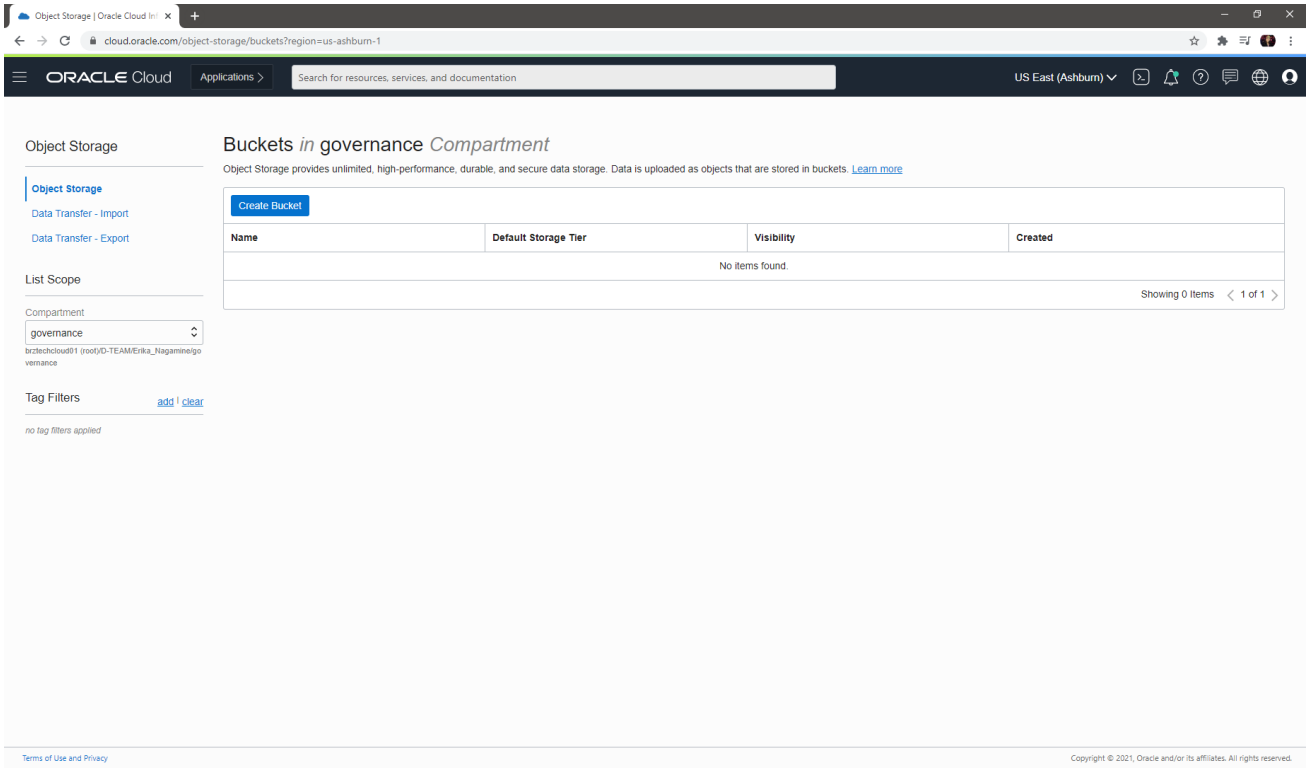
Configuração da máquina adicional para criação de funções

- É possível criar funções através de um computador Always Free, mas não cobriremos essa possibilidade no documento
- Não focaremos na visualização do dados que dentro do serviço de OAC (Oracle analytics Cloud), Oracle Machine Learning Notebooks presente no autonomous ou um relatório via APEX
- A infraestrutura completa também pode ser provisionada via infraestrutura como um código
- Temos por pressuposto que você possui acesso a cloud não cobriremos o tópico de fundação e boas práticas na cloud, como criação de compartments, etc.

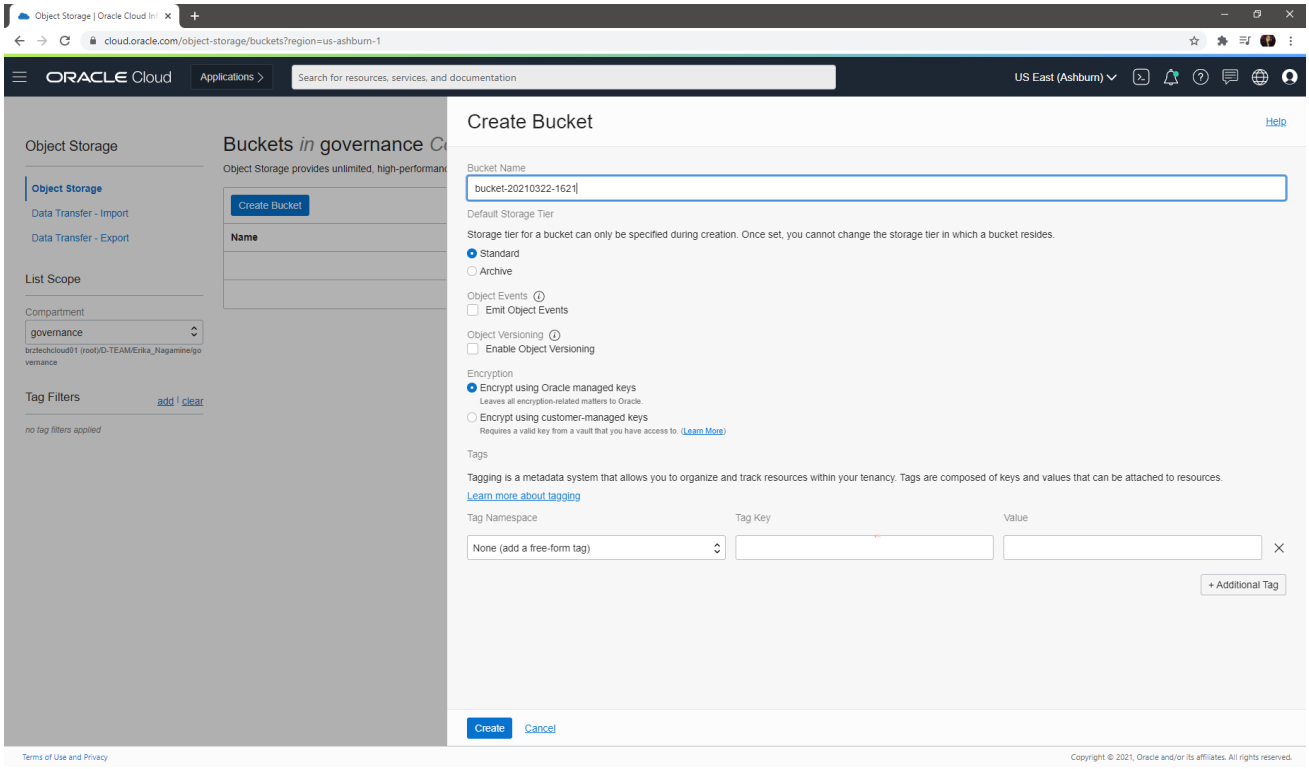
2. Provisionar os recursos

2.1.Object Storage

O object storage é um disco elástico para armazenamento dos mais diversos tipos de arquivos e cargas de trabalho: pode armazenar desde um dado não estruturado até um backup. Para criar um object storage, vá em menu -> Object storage -> object storage:



Iremos criar dois buckets onde um servirá para os dados processados e o outro para dados inseridos (onde vamos receber as informações. Na tela, clique na opção de “create bucket”.



Crie dois buckets com as seguintes informações:

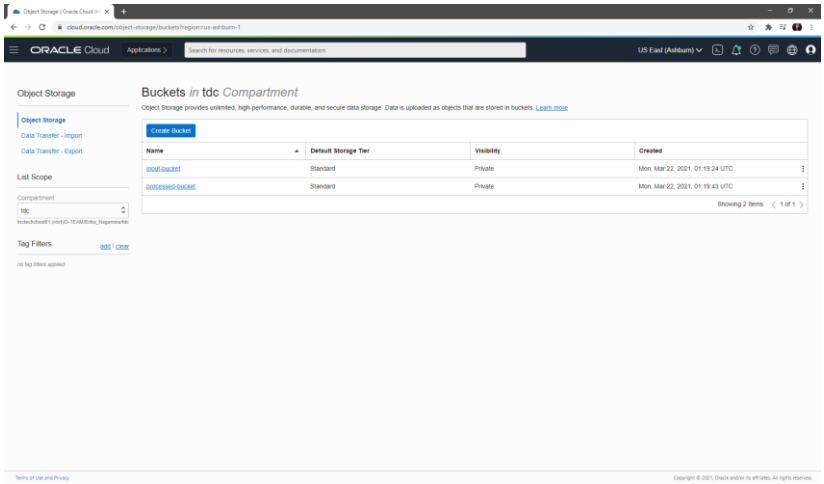
(bucket de entrada de dados)

- Bucket-name: input-bucket
- Object Events: Marque a opção para emitir eventos

(bucket de saida de dados)

- Bucket-name: processed-bucket
- Object Events: Marque a opção para emitir eventos

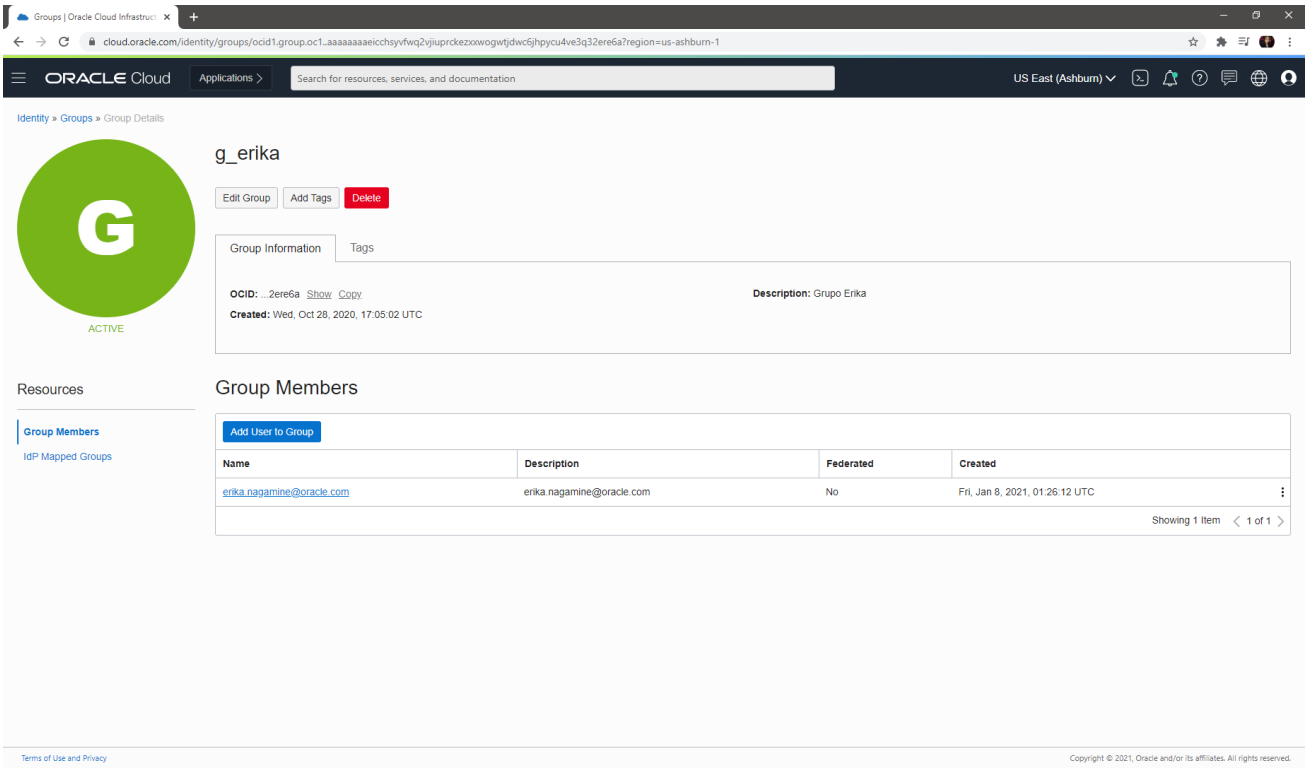
Ao final dos dois processos, você verá uma tela semelhante:



2.2. Criando funções

Alguns pré requisitos deverão ser realizados para esta fase:

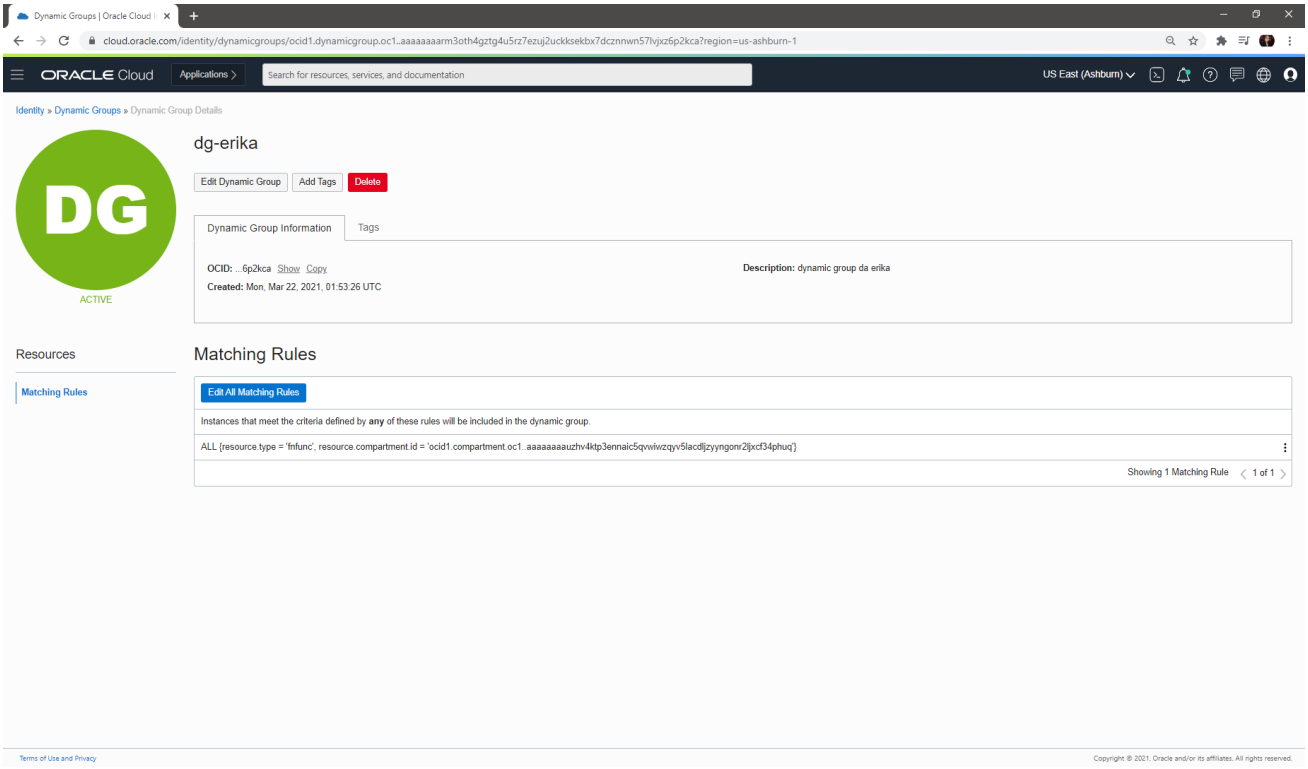
- Criação de grupos
  Em menu -> identity -> groups crie um grupo para atribuir as politicas necessárias a criação de funções. Um exemplo desta criação pode ser este:



  Atribua todos os usuários que precisam ter acesso a este grupo.

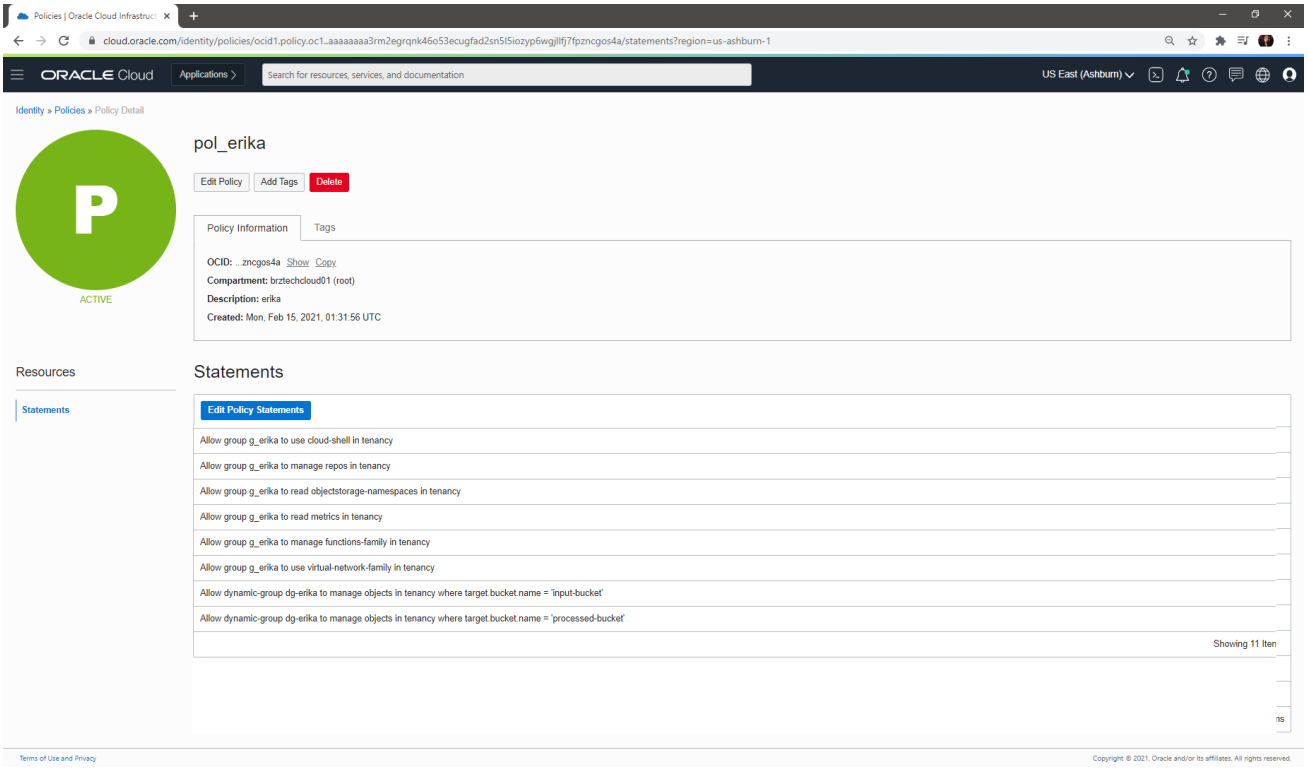
- Crie um grupo dinâmico
  Para atribuição de permissões sobre seu compartment especifico, utilize o recurso de grupo dinâmico. Para cria-lo vá em menu -> identity -> Dynamic Group

```
ALL {resource.type = 'fnfunc', resource.compartment.id =
'id_do_compartment'}
```

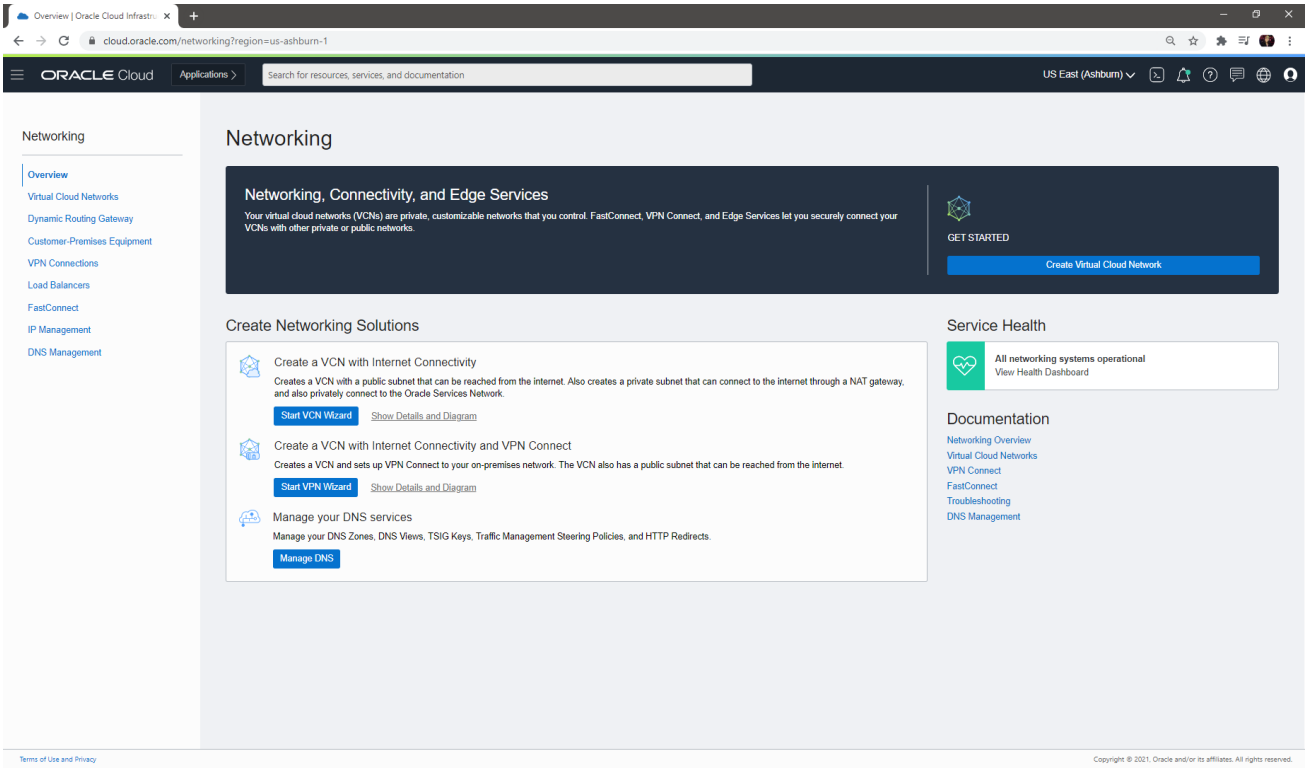


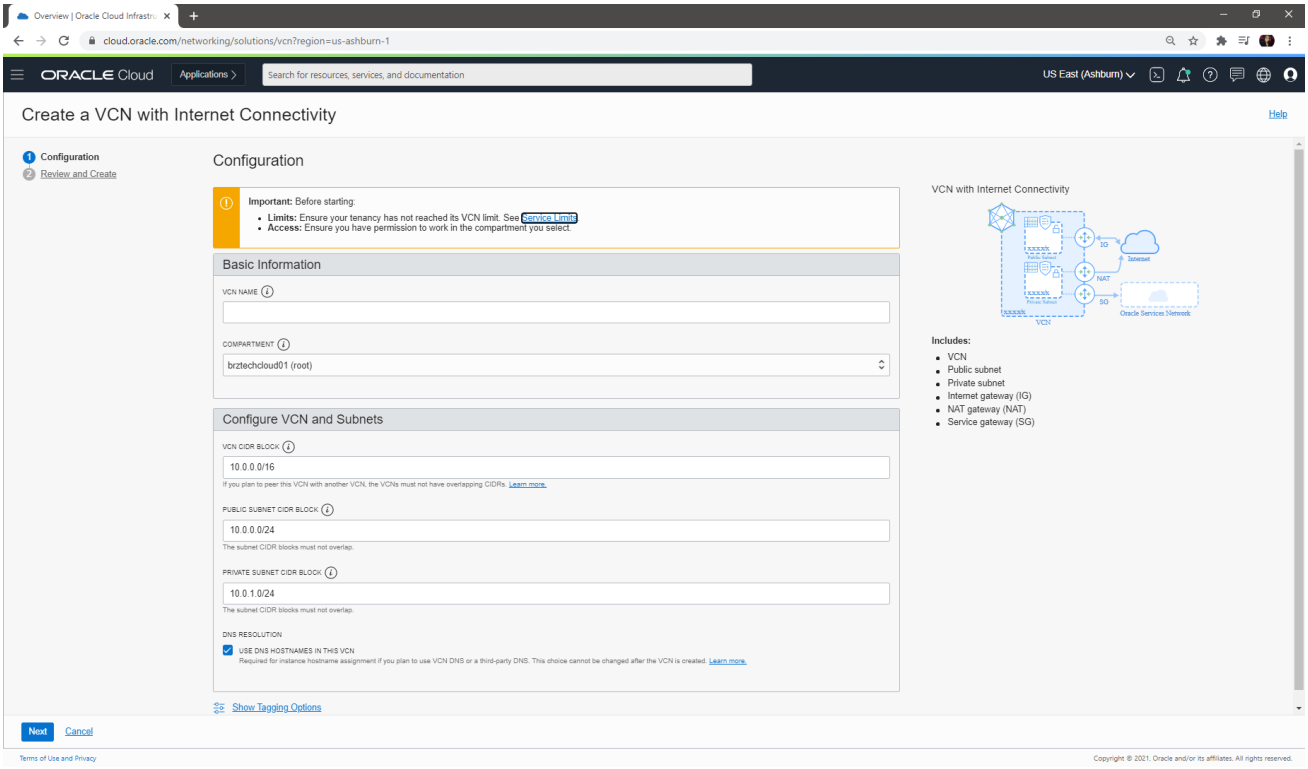
- Criação de políticas
  Em menu -> identity – Policies crie as politicas necessárias para esta fase. As politicas necessárias são descritas abaixo:

```
Allow group g_erika to use cloud-shell in tenancy

Allow group g_erika to manage repos in tenancy

Allow group g_erika to read objectstorage-namespaces in tenancy

Allow group g_erika to read metrics in tenancy

Allow group g_erika to manage functions-family in tenancy

Allow group g_erika to use virtual-network-family in tenancy

Allow dynamic-group dg-erika to manage objects in compartment <seu
compartment> where target.bucket.name=input-bucket

Allow dynamic-group dg-erika to manage objects in compartment <seu
compartment> where target.bucket.name=processed-bucket
```

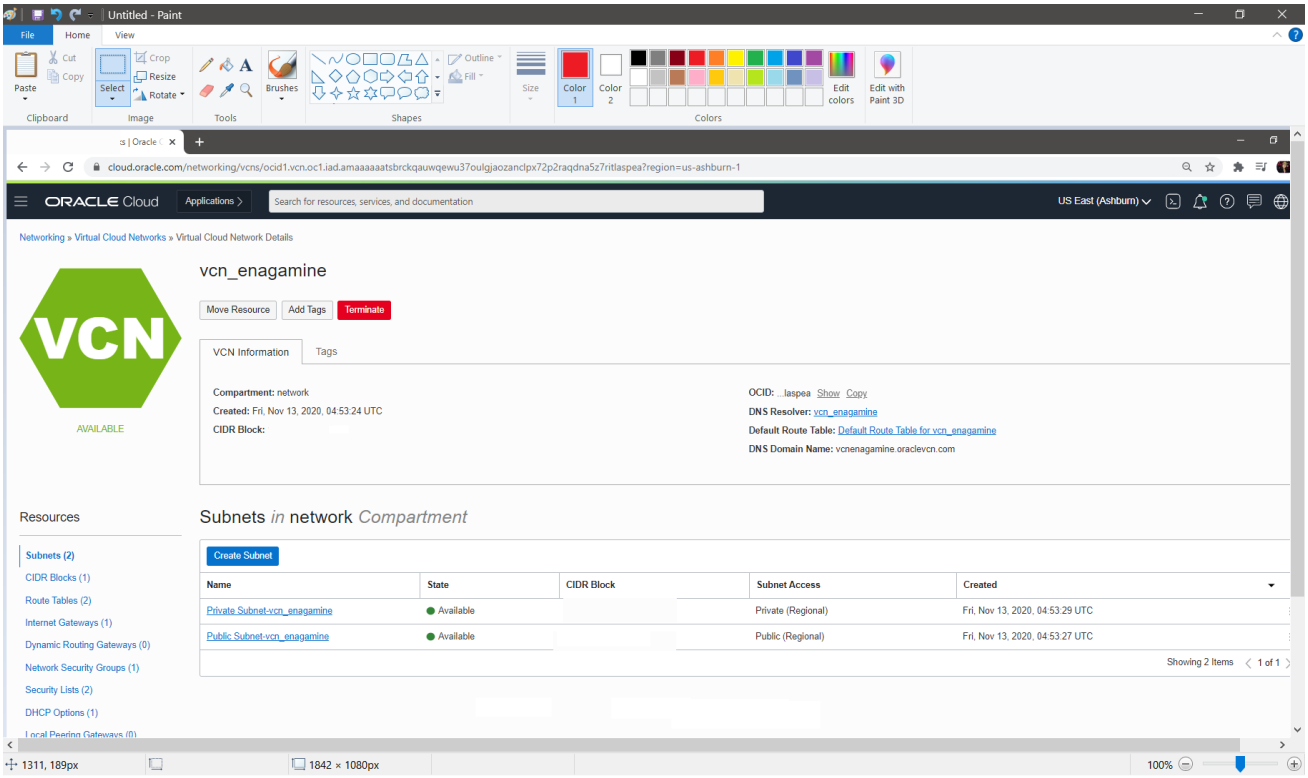


- Criação de VCN
  Neste documento criaremos uma rede padrão via “wizard” mas como padrão e boa pratica, pense sempre na boa fundação do seus recursos em cloud. Para criar a rede padrão, vá em menu -> networking -> start VCN Wizard e siga o passo a passo:

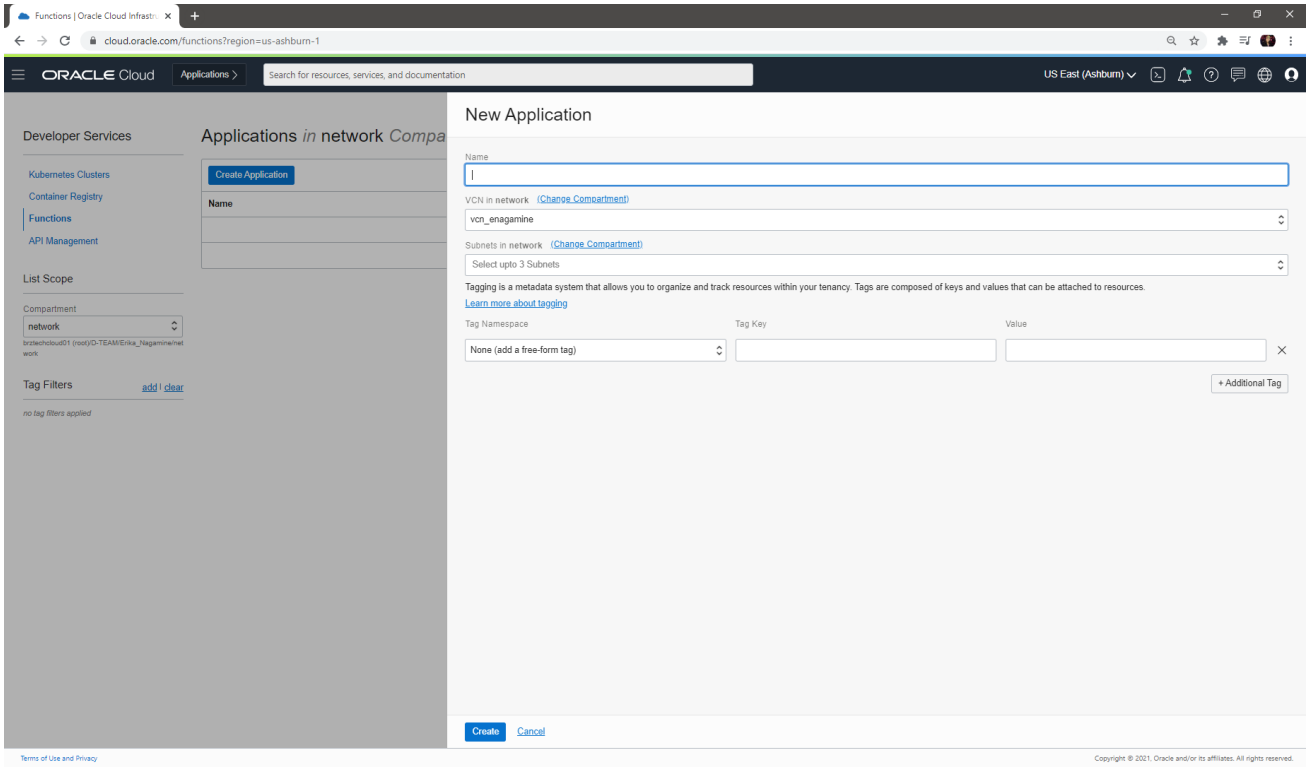




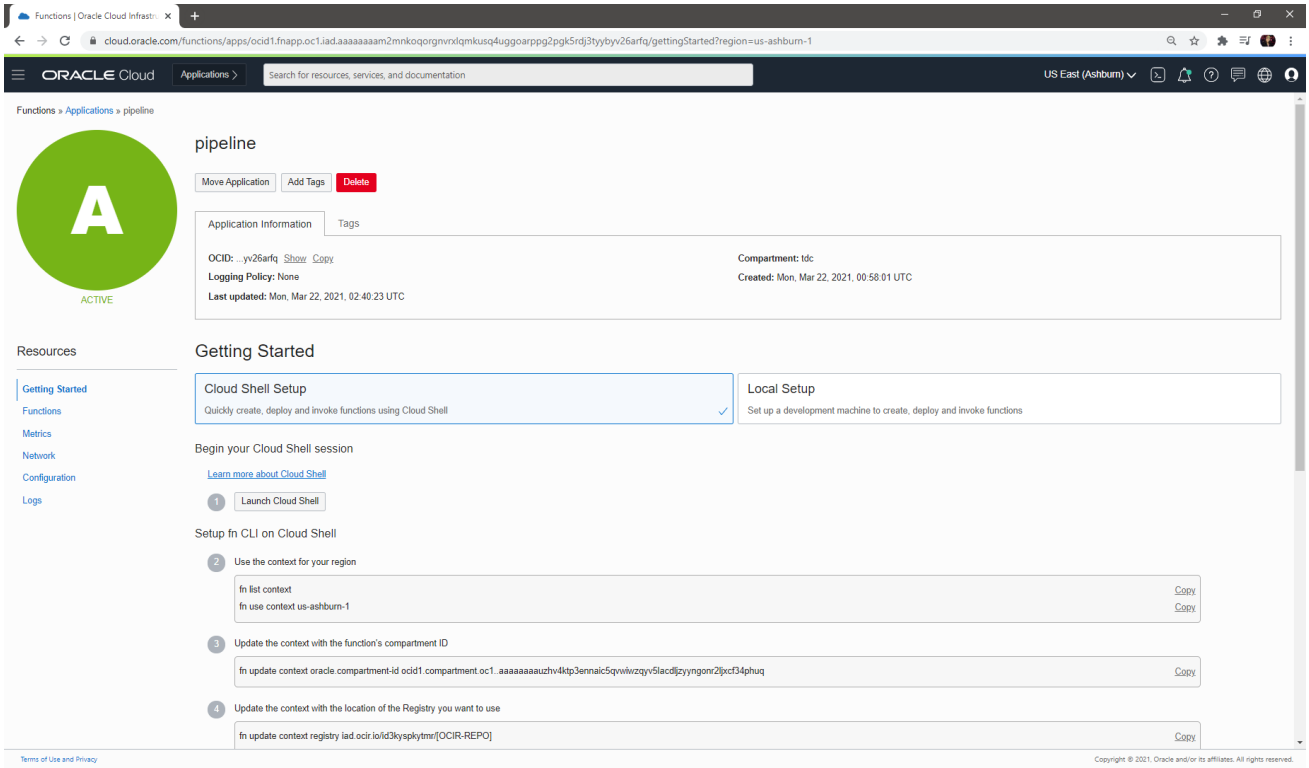
Ao final do processo você terá duas subnets:



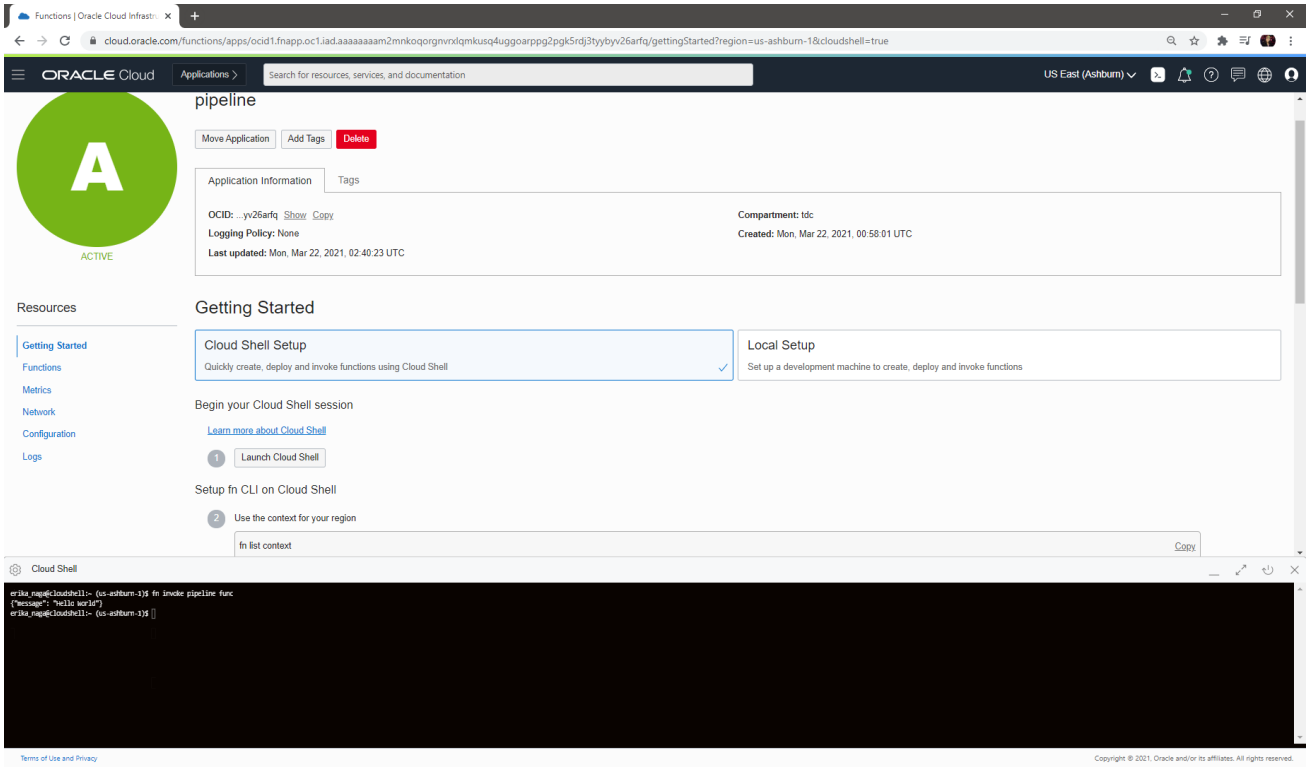
Agora estamos pronto para o processo de criação de funções. Para isso, vá em menu -> developer services -> functions e clique em create application:



Você precisará informar sua VCN criada. Sugerimos (para efeito de workshop, não de dia a dia, colocar na subnet publica).



Após criar sua aplicação para a função, agora vamos configurá-la. Para isso faremos o setup via cloud shell. Siga o passo a passo da página e não se esqueça de gerar e anotar o token:



Agora que a configuração básica da sua função está realizada, no cloud shell baixe os scripts abaixo e coloque os no diretório func:

```
wget https://raw.githubusercontent.com/oracle/oracle-functions-samples/master/samples/oci-load-file-into-adw-python/func.py

wget https://raw.githubusercontent.com/oracle/oracle-functions-samples/master/samples/oci-load-file-into-adw-python/requirements.txt

wget https://raw.githubusercontent.com/oracle/oracle-functions-samples/master/samples/oci-load-file-into-adw-python/func.yaml
```

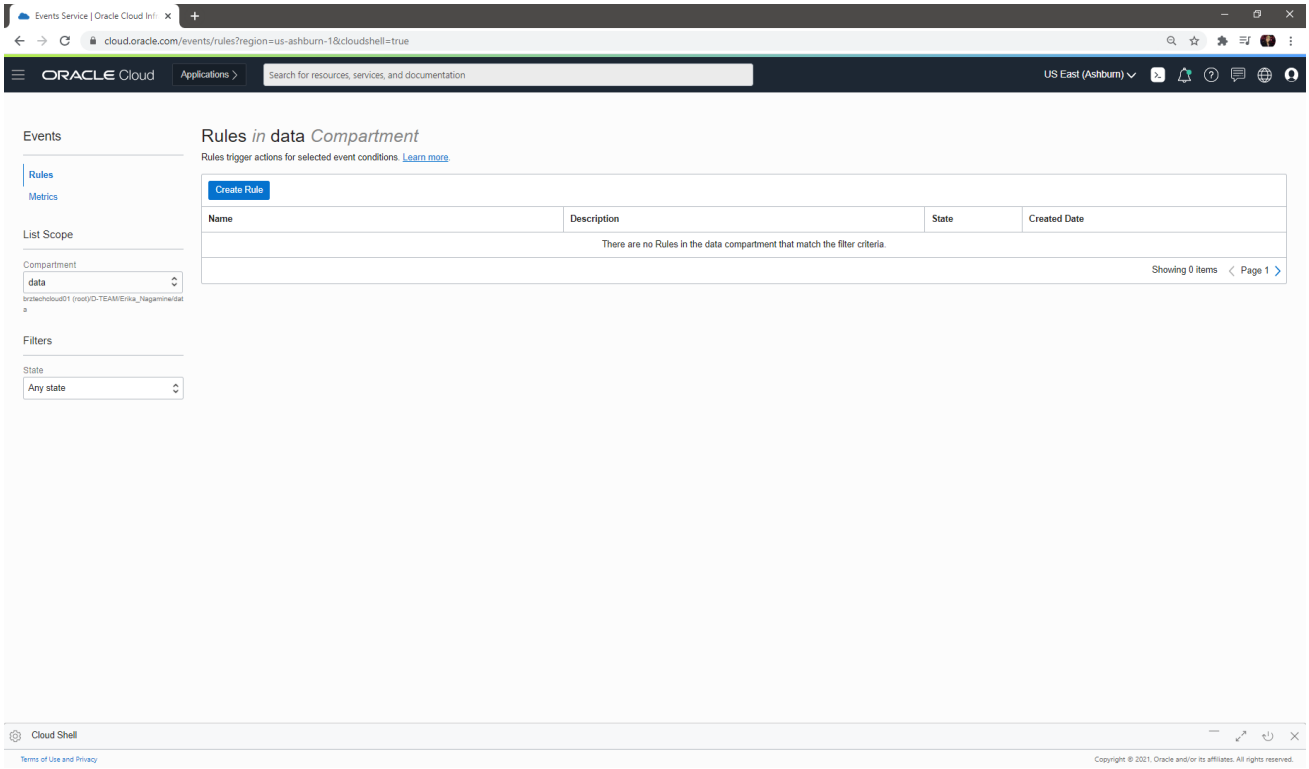
Faça o deploy da sua função com novas parametrizações:

```
fn -v deploy --app pipeline func
```

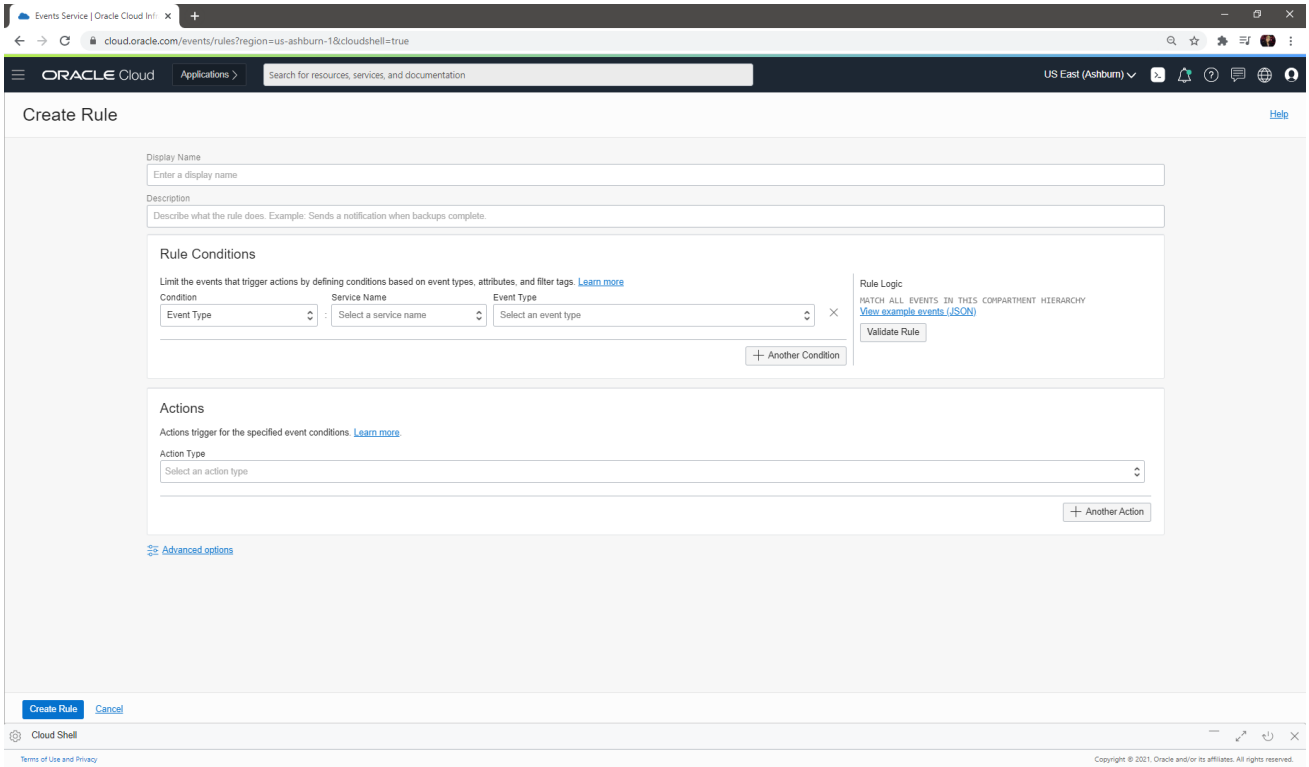
Retornaremos a este passo posteriormente para configurar os parâmetros que sua função irá receber.

2.3. Criando uma evento

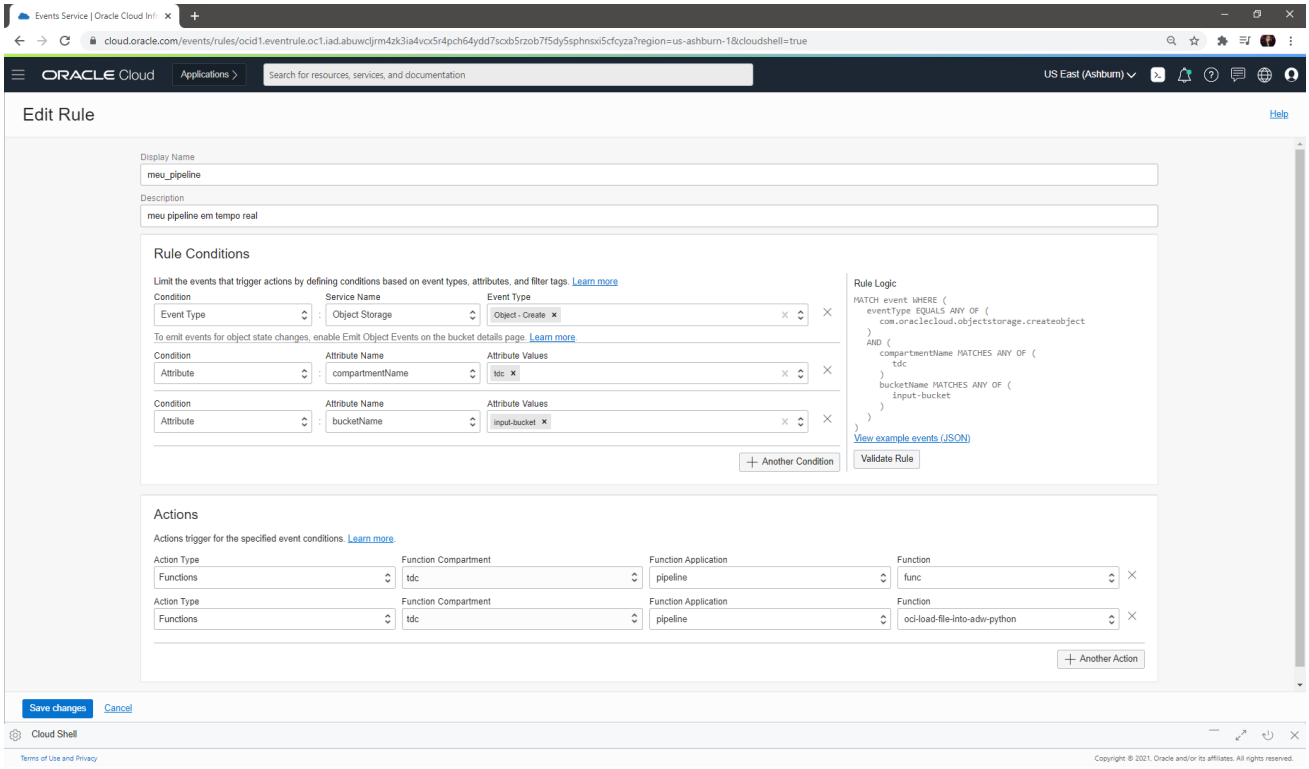
Navegue até menu -> application integration -> event service



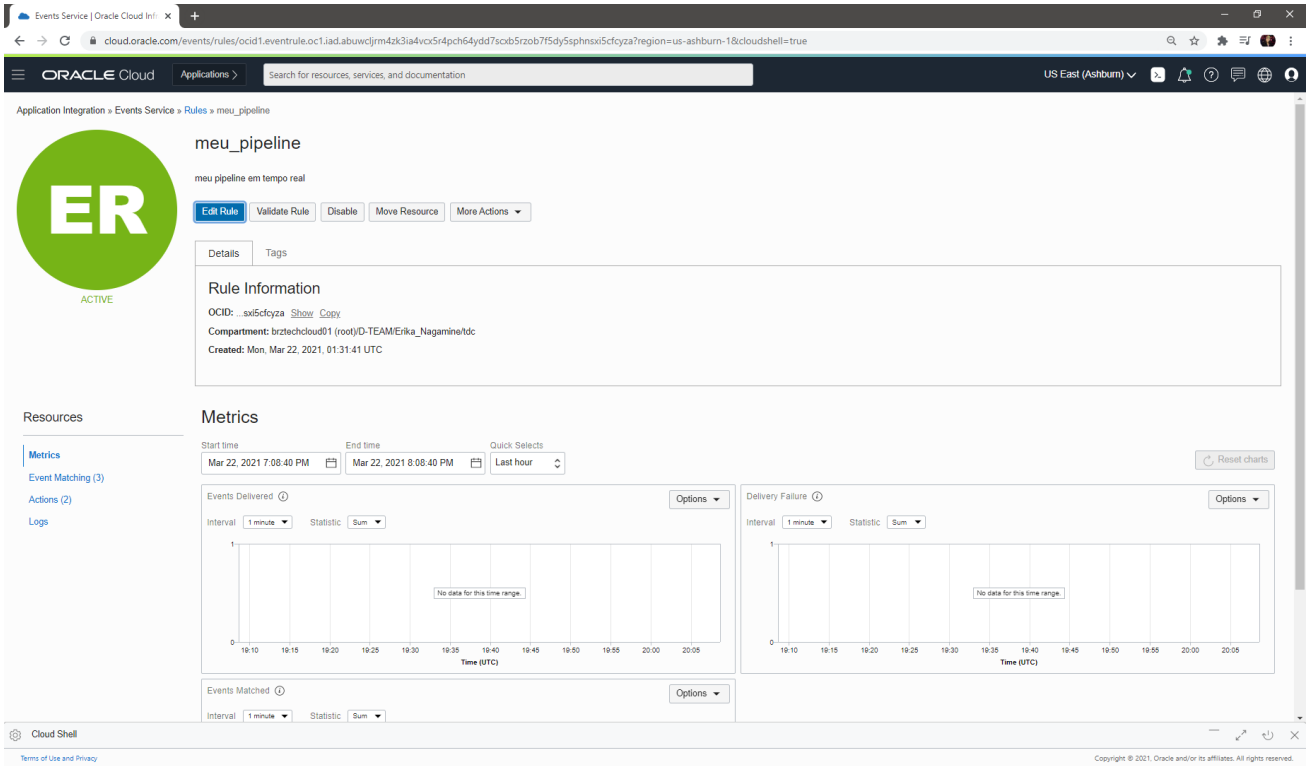
Clique em criar uma nova regra:



Utilize as parametrizações abaixo:



Salve as configurações:

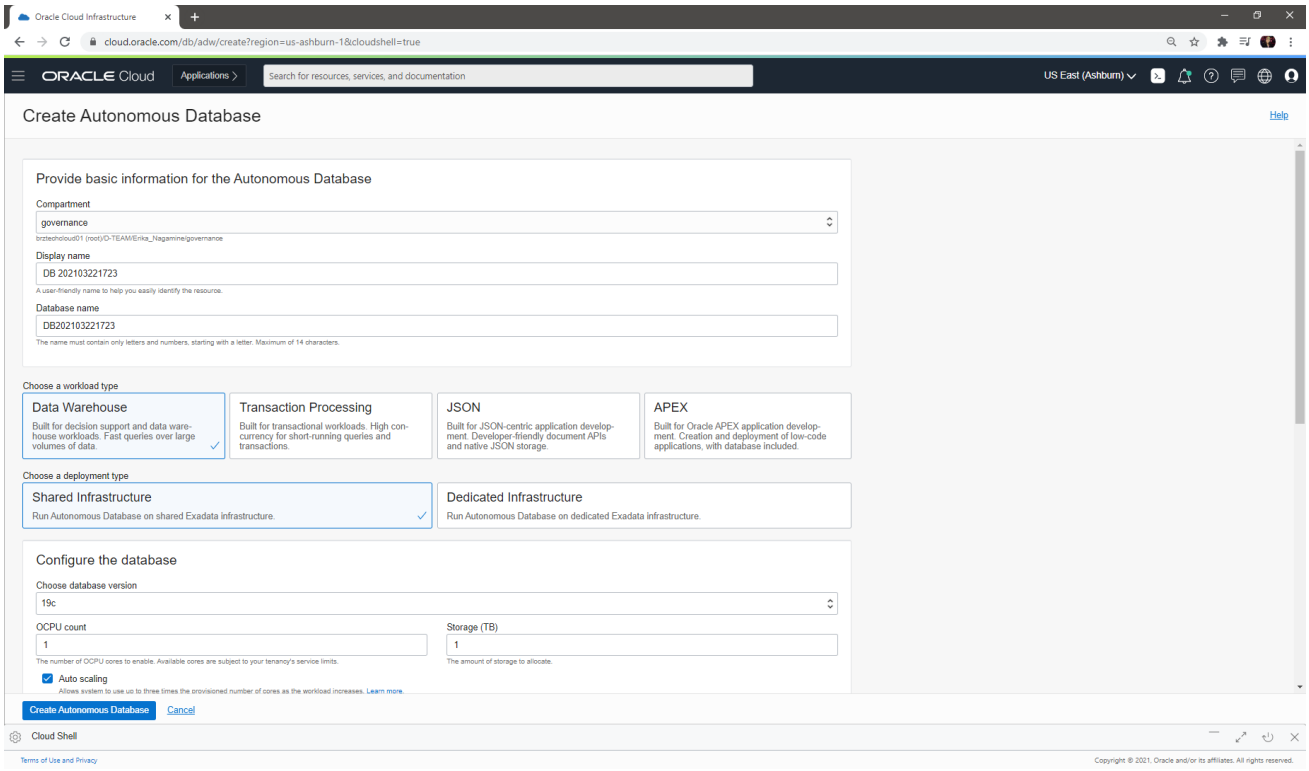


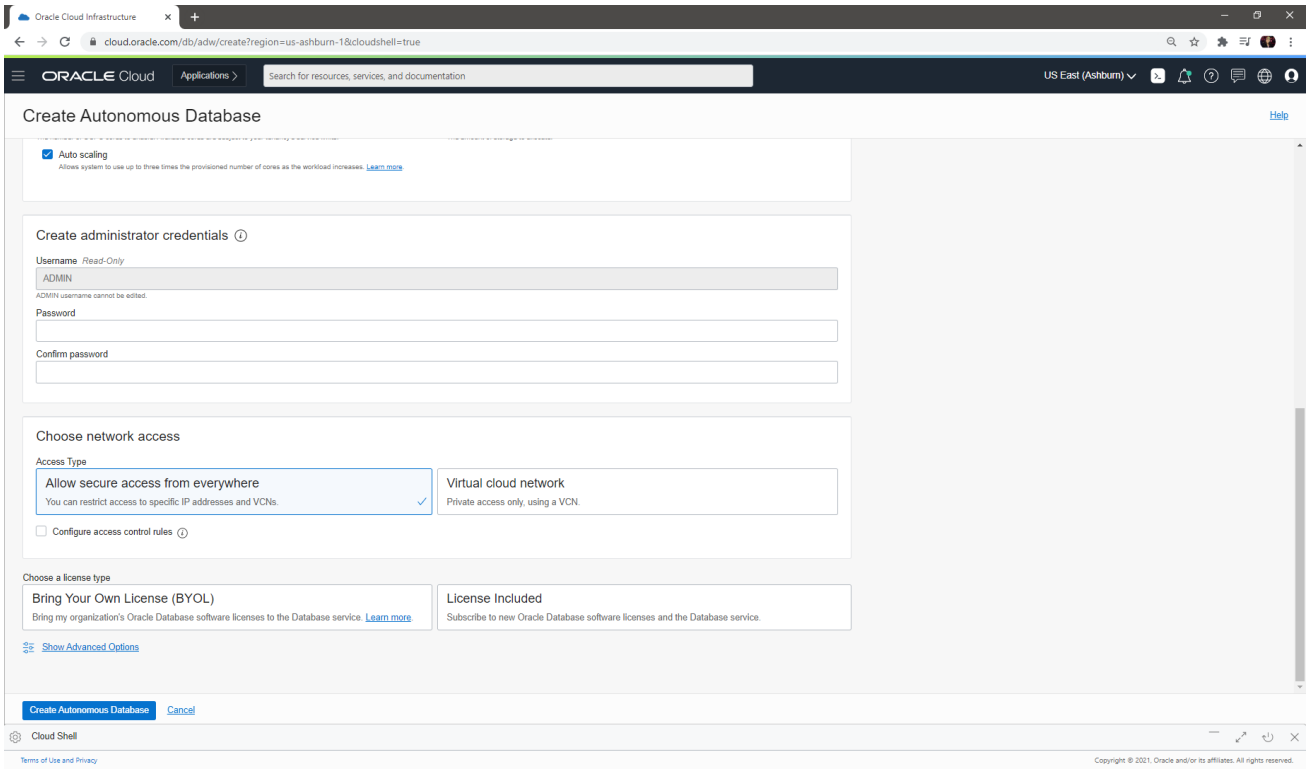
Nesse momento seu data lake está pronto para receber os dados. Faremos a configuração do nosso data warehouse

2.4. Banco de dados Autonomous

Nosso Data warehouse será construído em cima do autonomous database. Navegue no menu direito até autonomous database e clique em criar autonomous data warehouse.

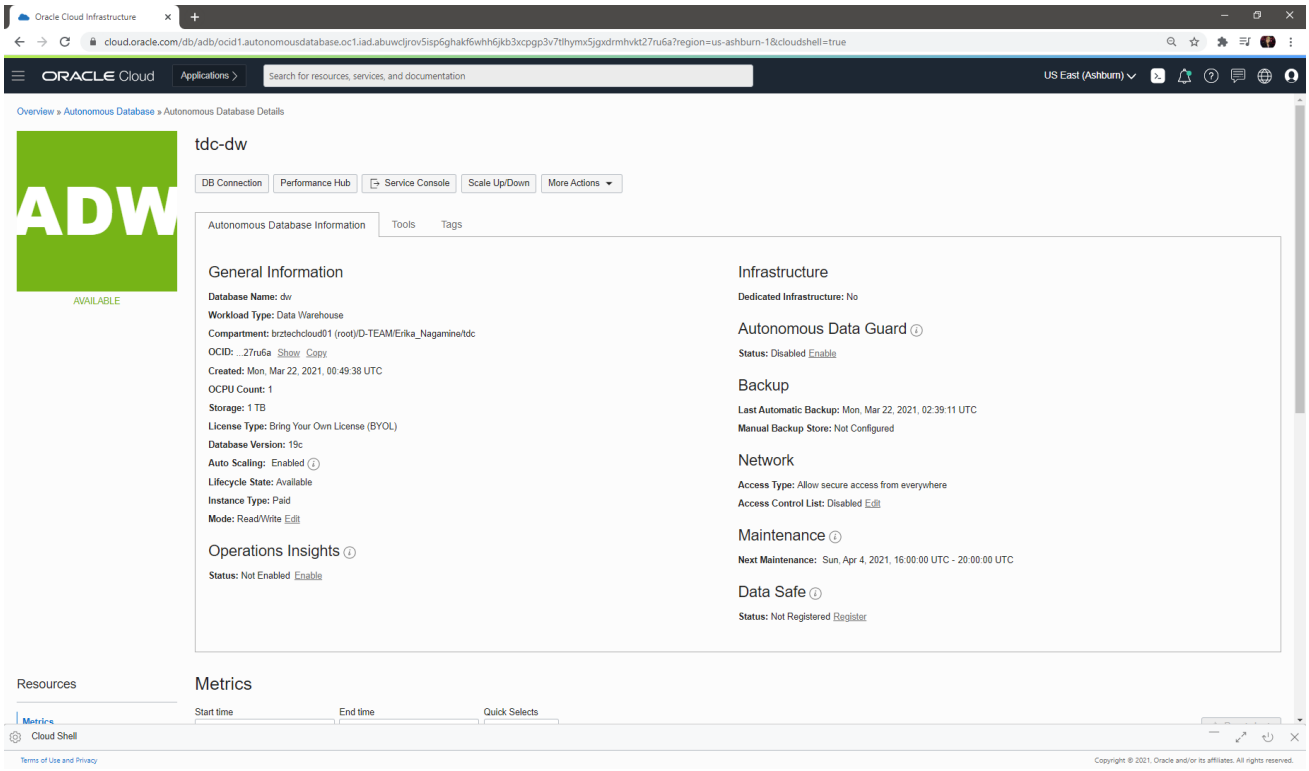
Em autonomous data warehouse clique em “create autonomous database:





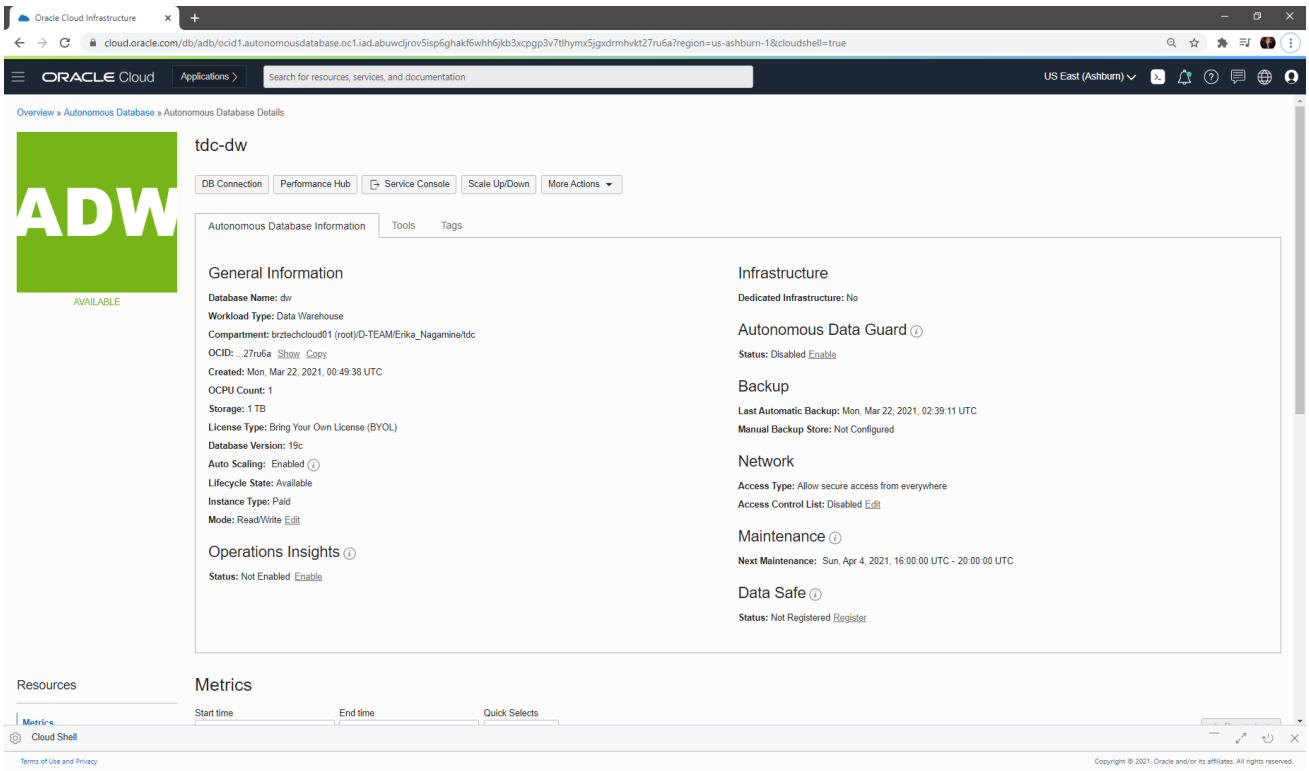
Preencha com as seguintes informações (note que é possível provisionar o autonomous como Always free:

- Compartment: se você criou um compartment, selecione o responsável:
- Display name: tdc-dw
- Database name: dw
- Database version: 19c
- Workload type: data warehouse
- User password: coloque uma de sua escolha
- License type: Byol (se você selecionou a opção Always free do autonomous desconsidere essa parte)

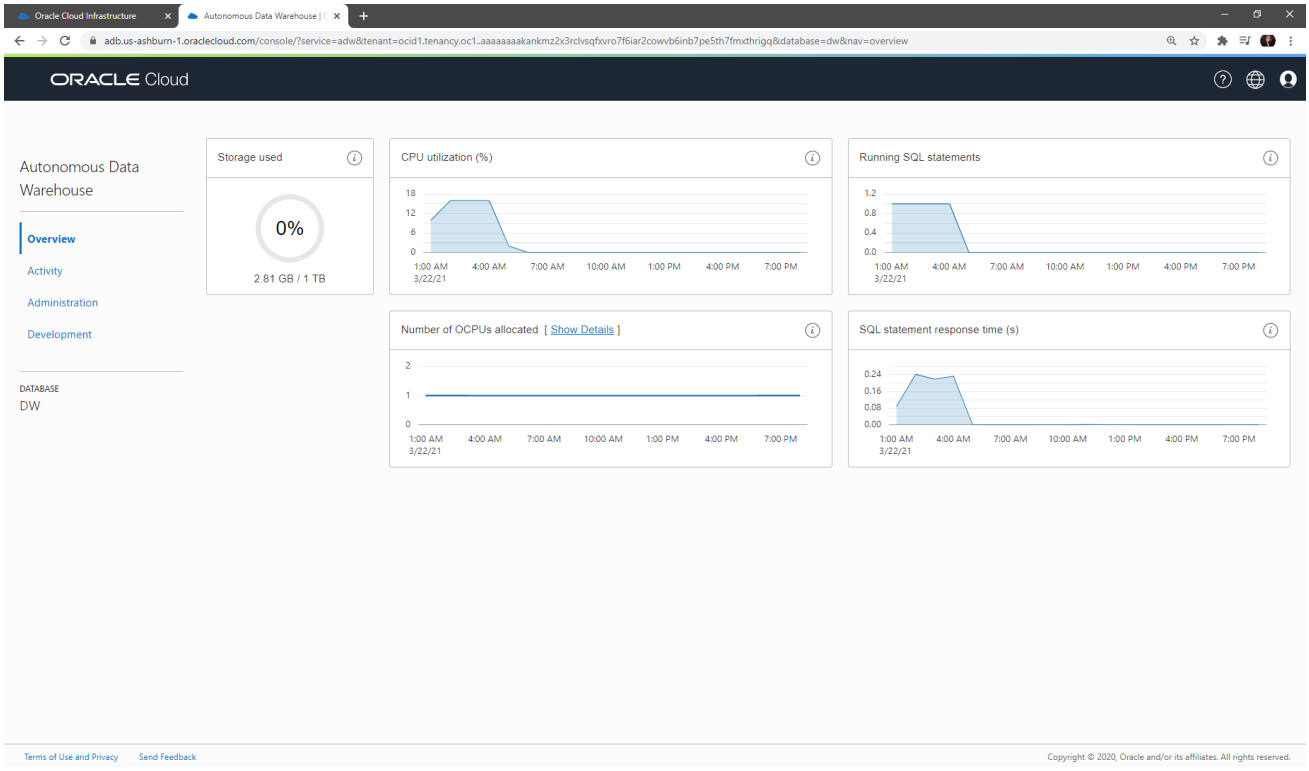


3. Configurando seu ambiente

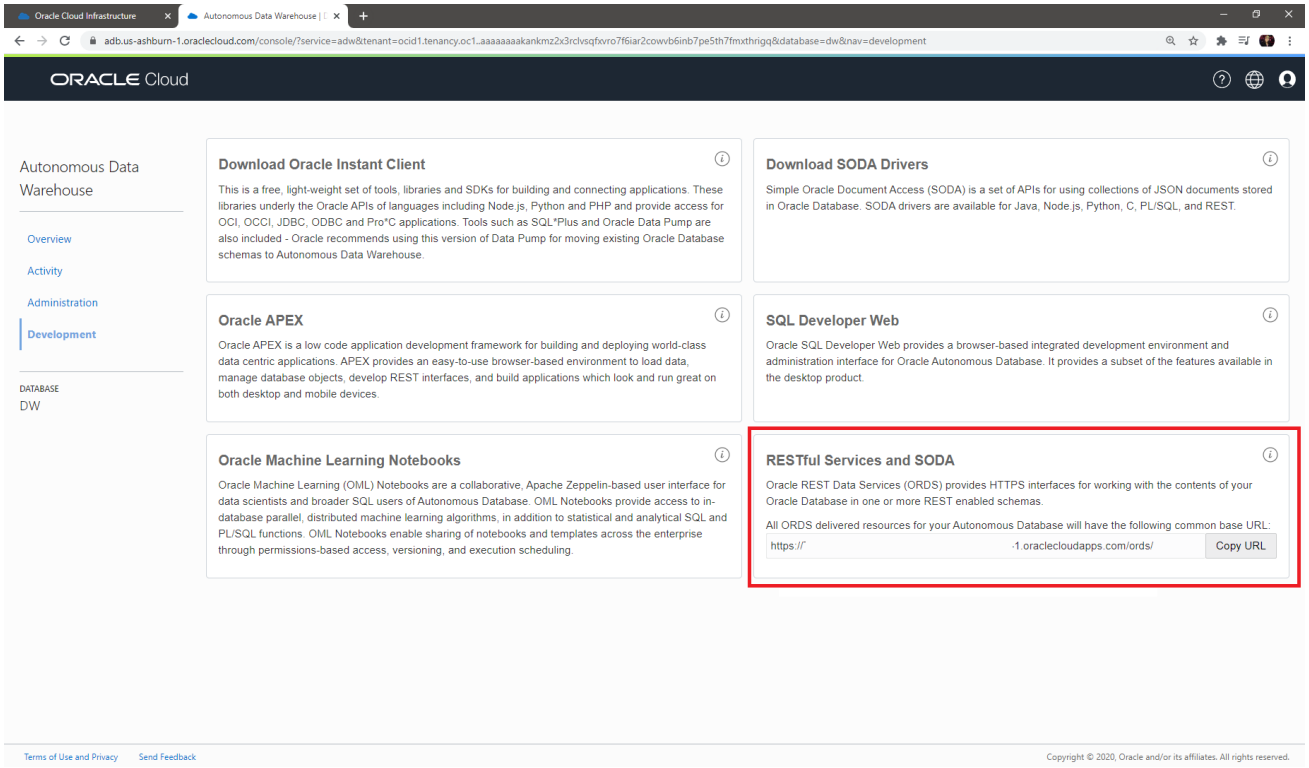
Agora que o stack está provisionado, vamos integrar os recursos. O primeiro passo é validar sua conexão via API para o Autonomous database e criar uma colection. Vamos fazer isso tudo através de linha de comando no cloud shell. Vamos pegar a informação necessária para este step, vá no autonomous criado e clique em service console:



Dentro da service console, navegue em development:



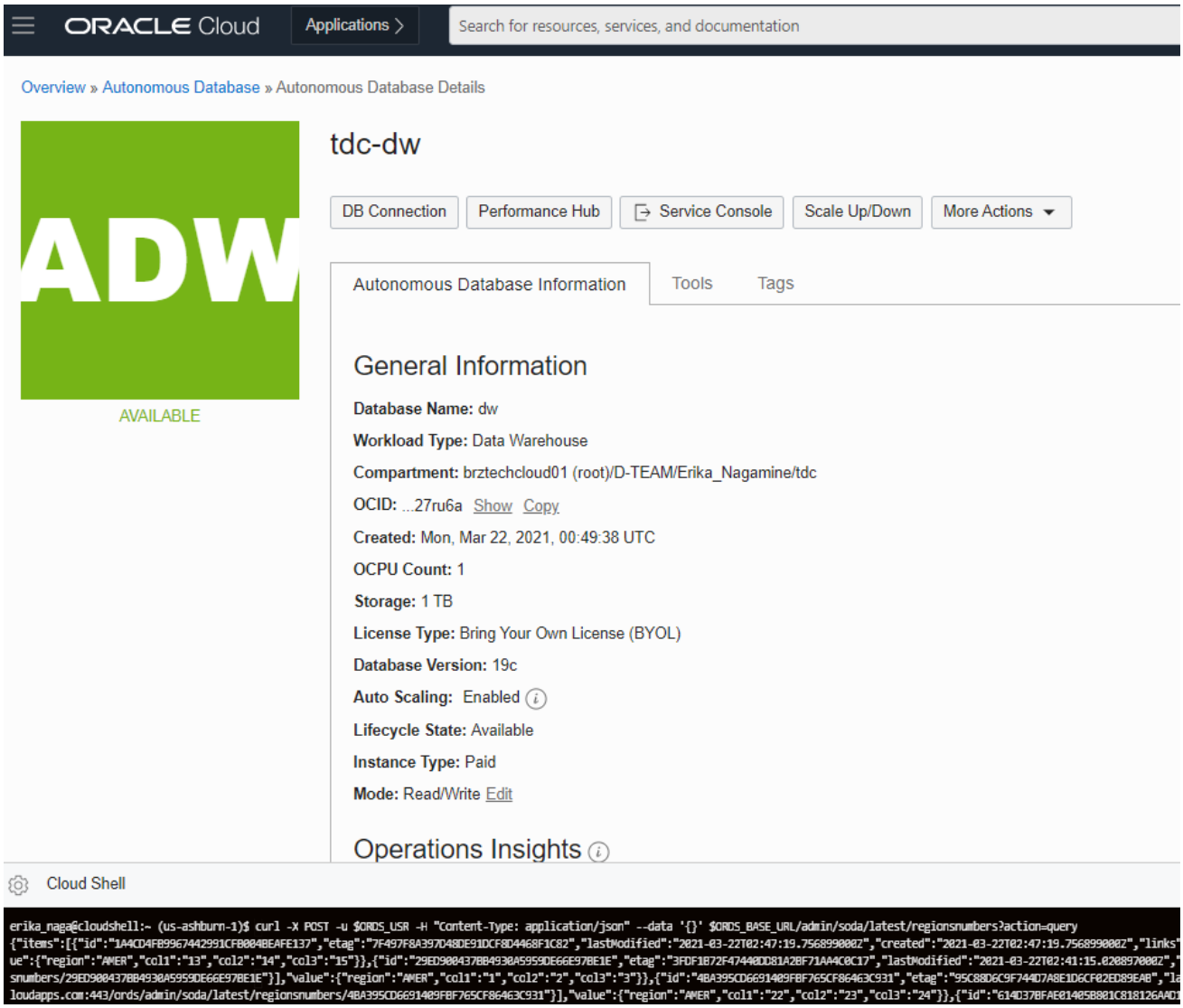
Dentro de development, copie a URL do RESTful services and SODA:



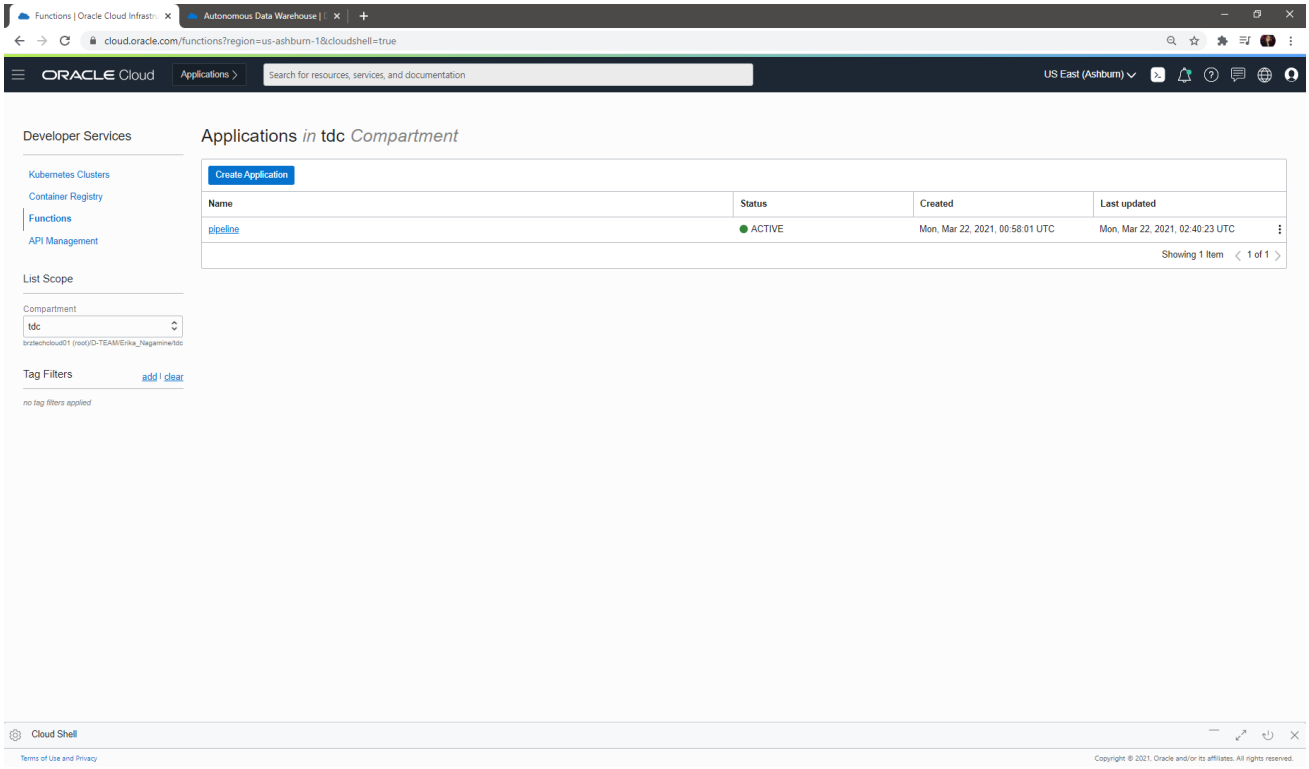
Agora substitua as variaveis abaixo e execute no seu

```
export ORDS_BASE_URL=< URL do SEU ADW – Restful and SODA>
export ORDS_USR='<usuario>:<senha>'
```

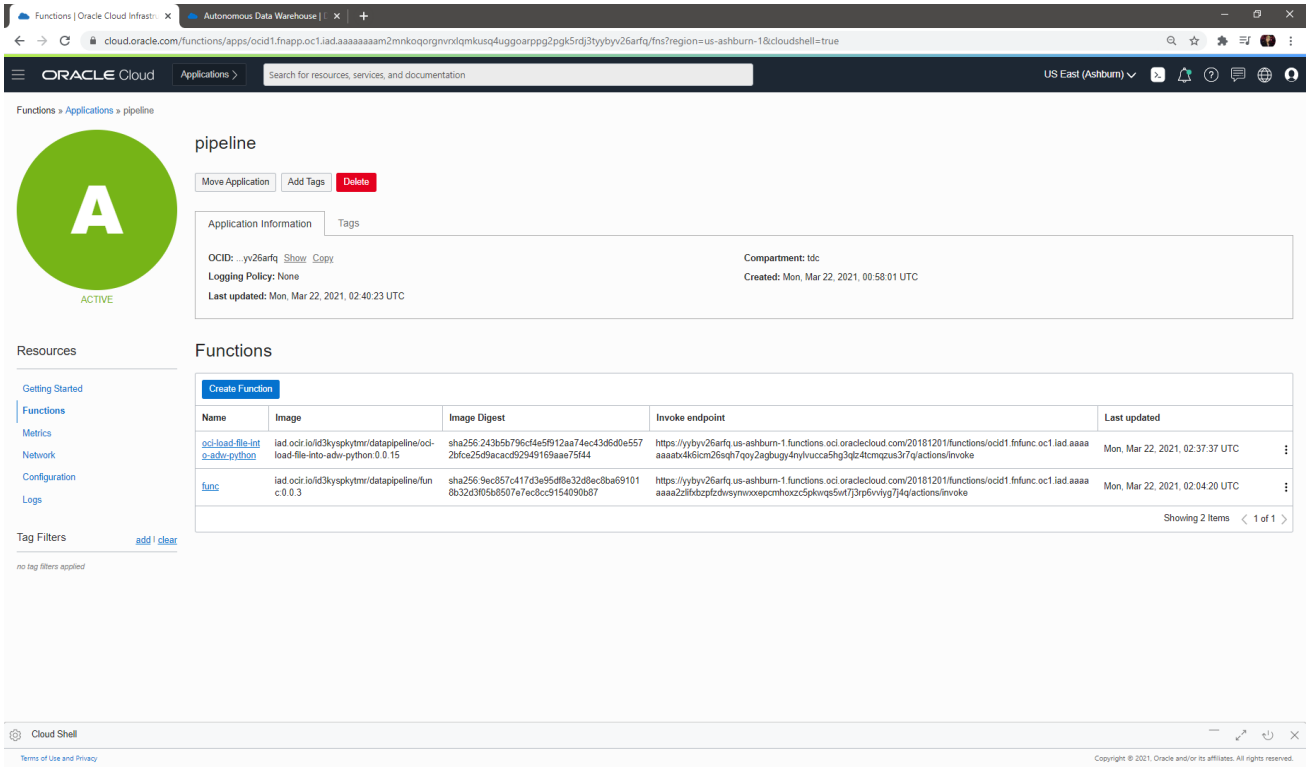
```
curl -X PUT -u $ORDS_USR -H "Content-Type: application/json"
$ORDS_BASE_URL/admin/soda/latest/regionsnumbers

curl -u $ORDS_USR -H "Content-Type: application/json" $ORDS_BASE_URL/admin/soda/latest/

curl -X POST -u $ORDS_USR -H "Content-Type: application/json" --data '{}'
$ORDS_BASE_URL/admin/soda/latest/regionsnumbers?action=query

curl -X POST -u $ORDS_USR -H "Content-Type: application/json"
$ORDS_BASE_URL/admin/soda/latest/regionsnumbers?action=truncate
```



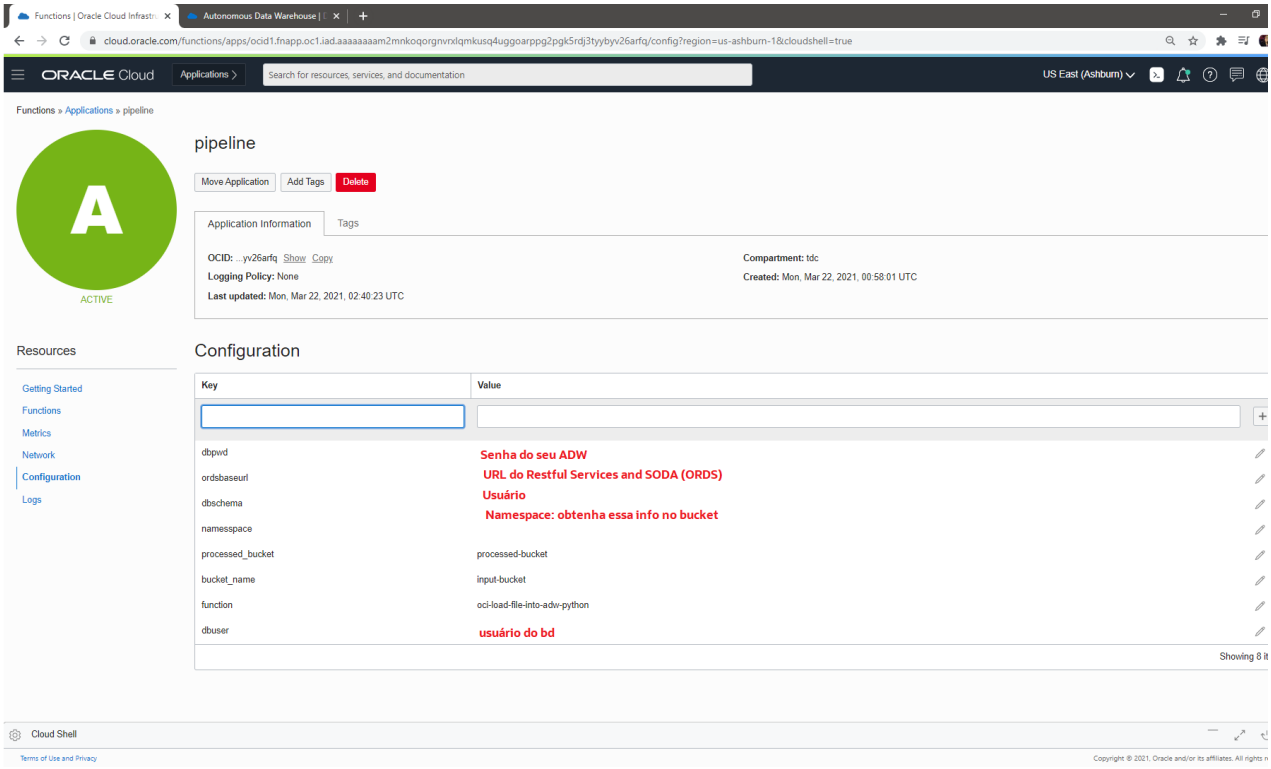
Mediante a esse funcionamento, o próximo passo é configurar a parametrização básica da sua função. Para isso vá em developer services -> functions:



Clique na aplicação criada:



Vá em configuração e sete as seguintes parametrizações:

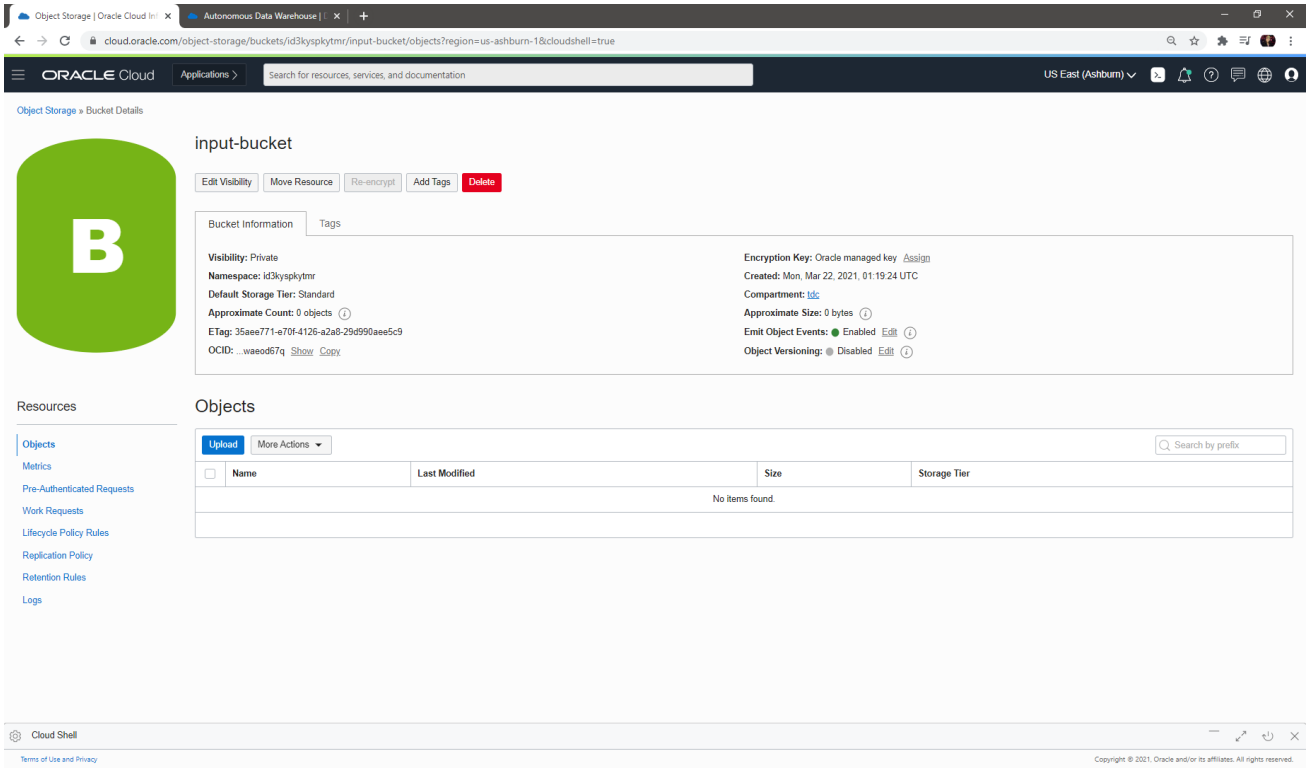


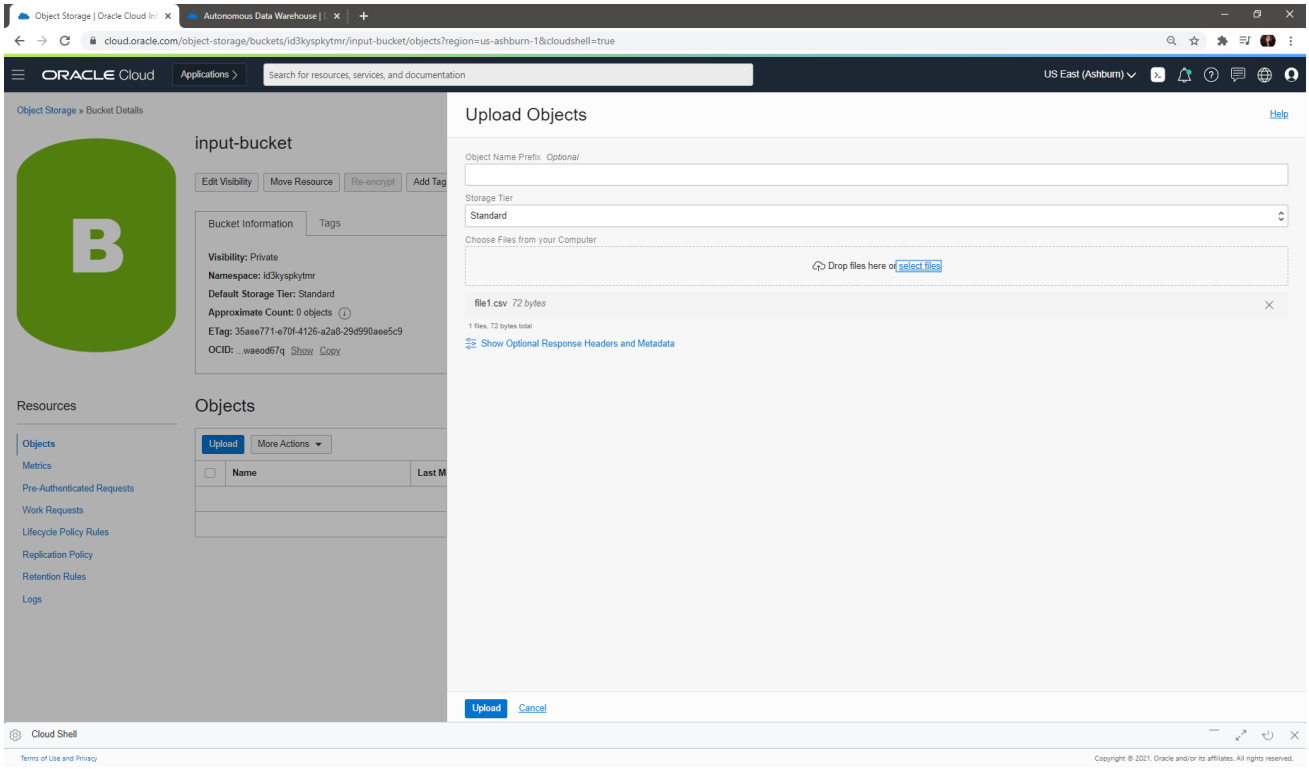
Pronto! Agora você está pronto para testar seu pipeline! 😊

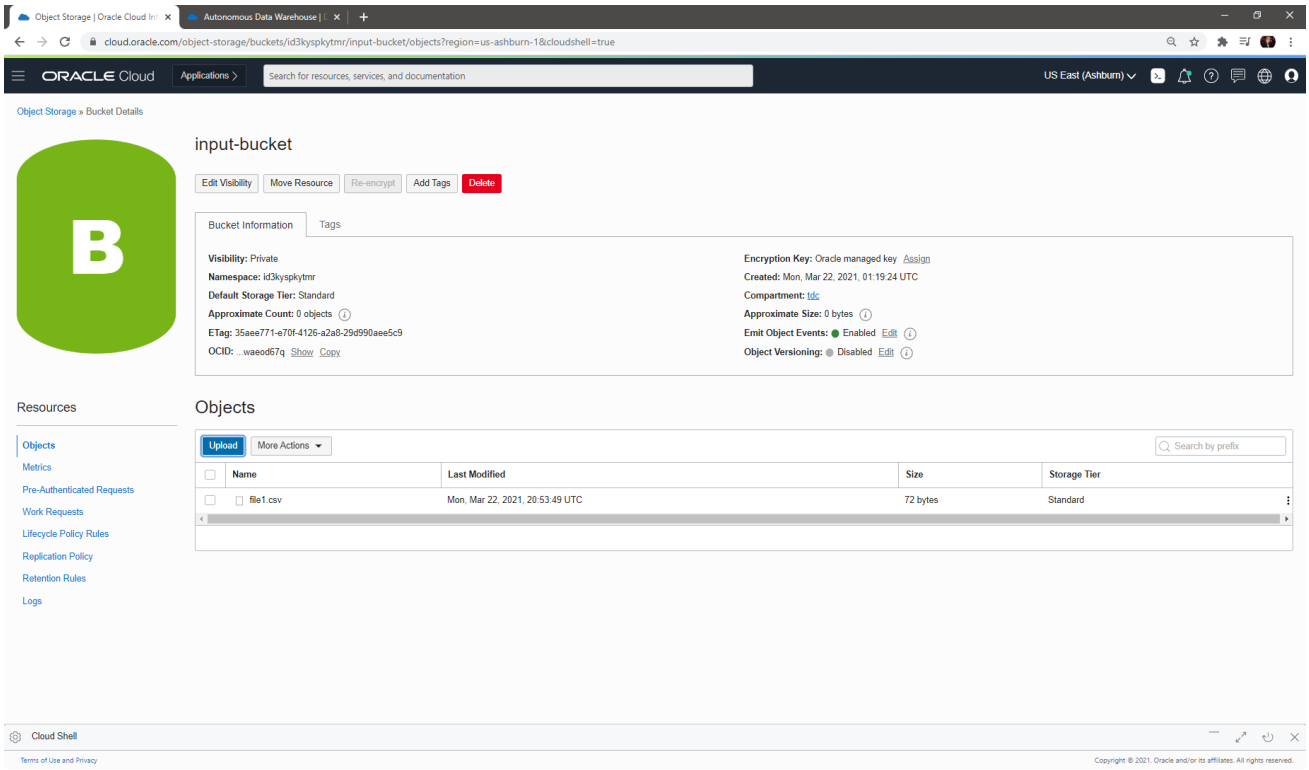
4. Testando seu pipeline de dados

Agora que o stack está todo configurado, é só colocar um arquivo no bucket input-bucket. Os arquivos de exemplo podem ser baixados do github:

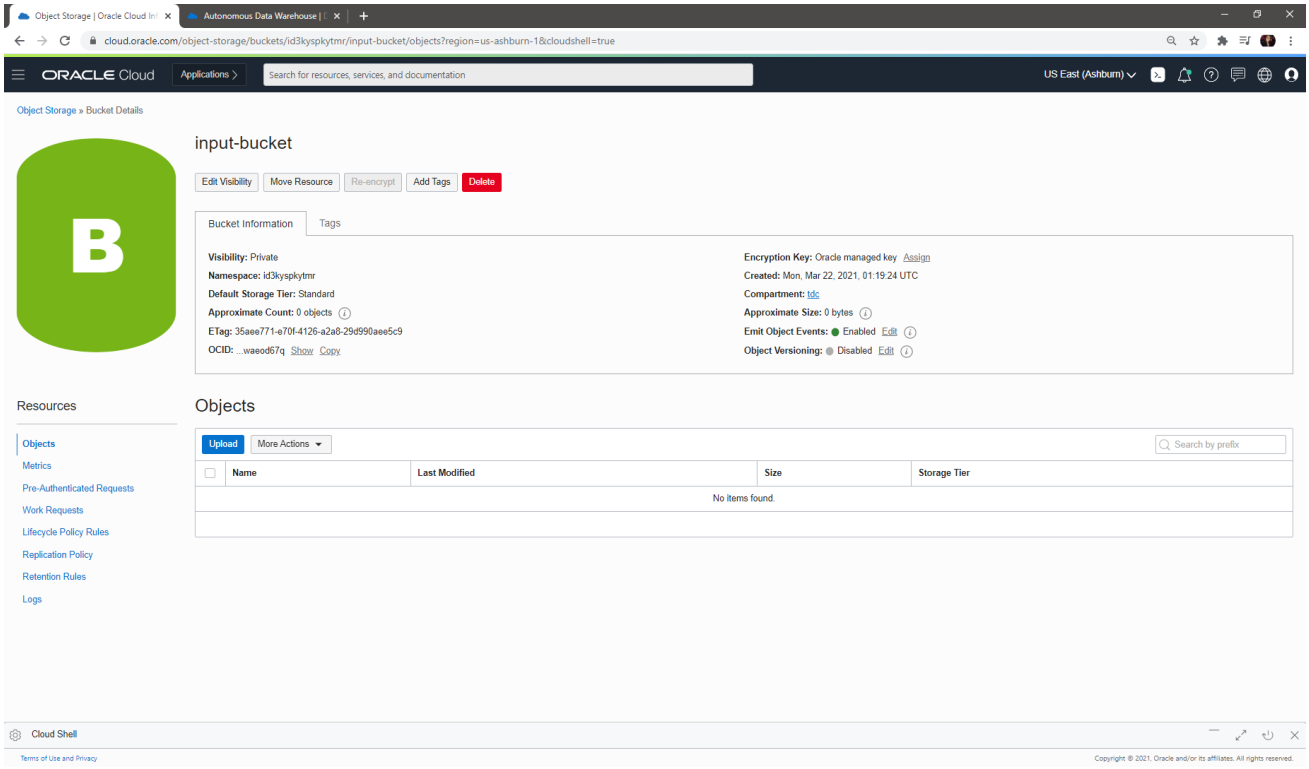
http://



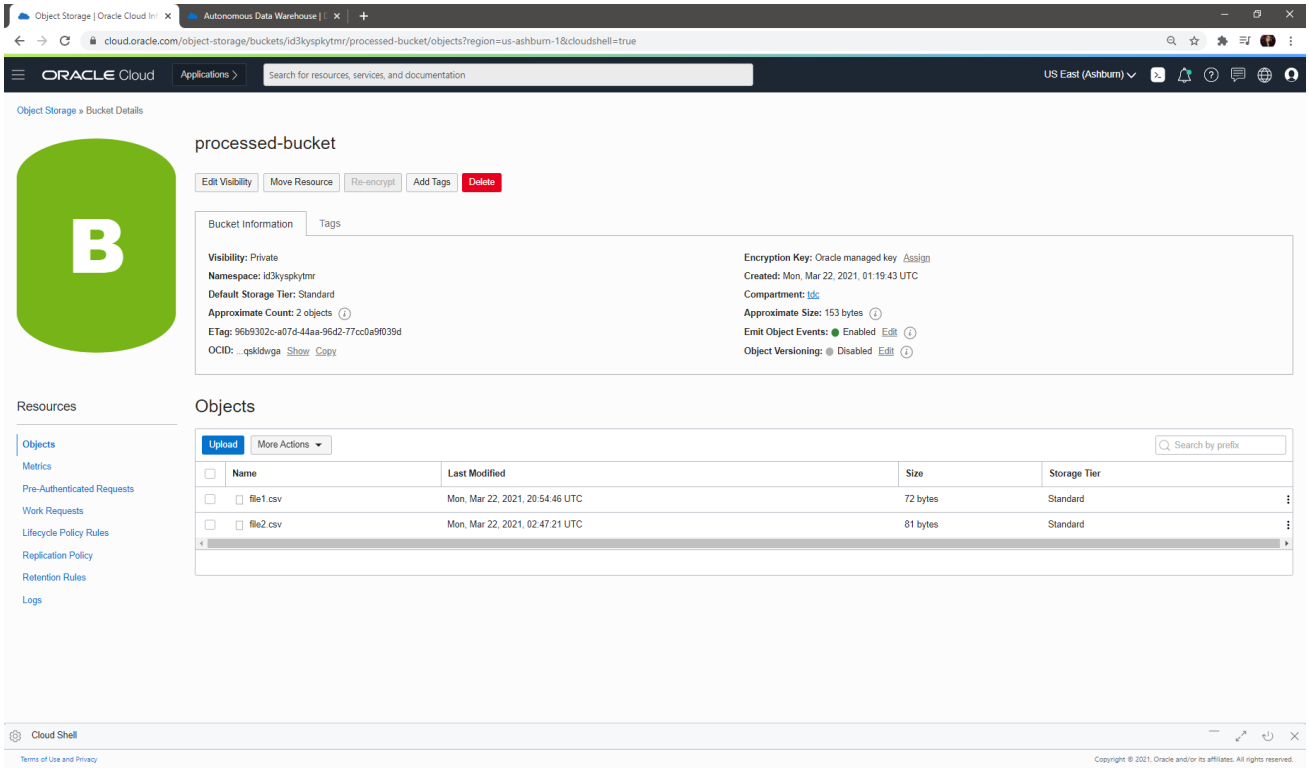




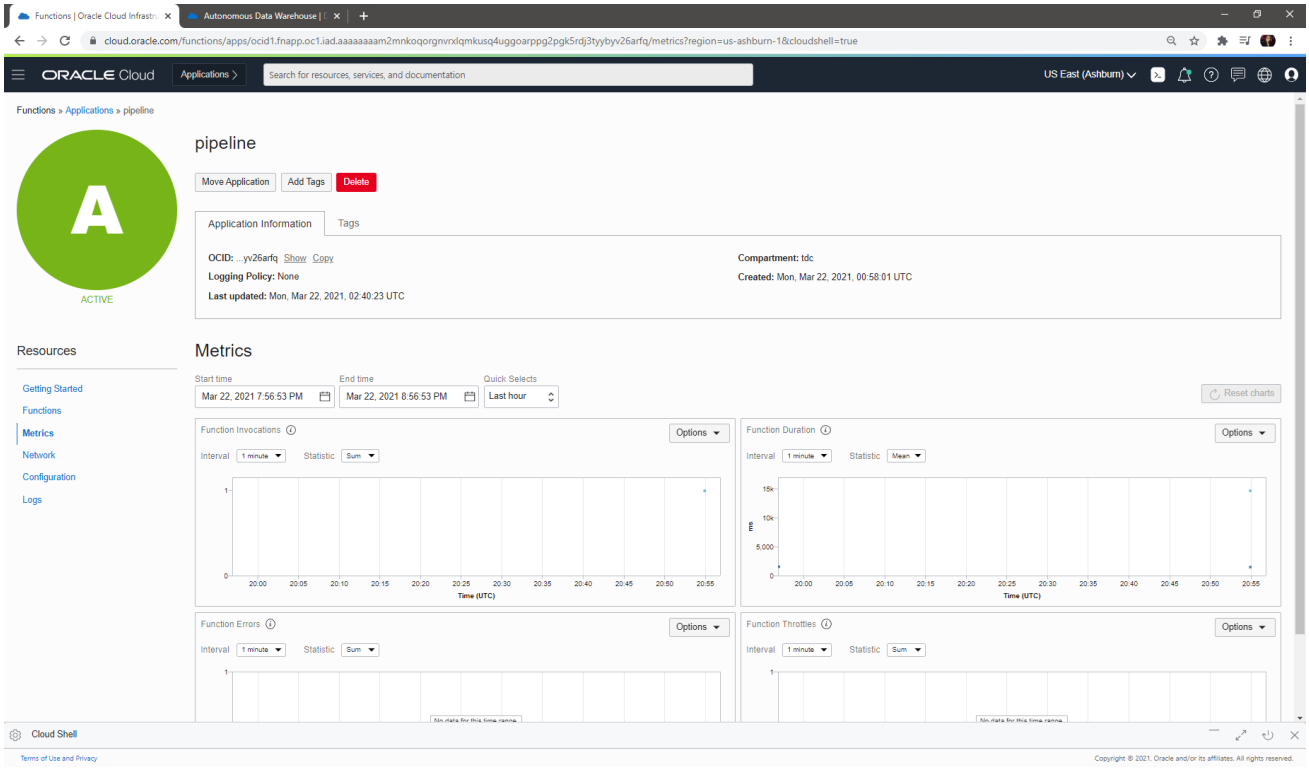
Aguarde um momento até o processo ser concluído. Quando ele for, veja que o arquivo não estará mais presente no bucket input-bucket – o tempo do processo depende do tamanho do arquivo (se for os de exemplo, note que será menos de 1 min):



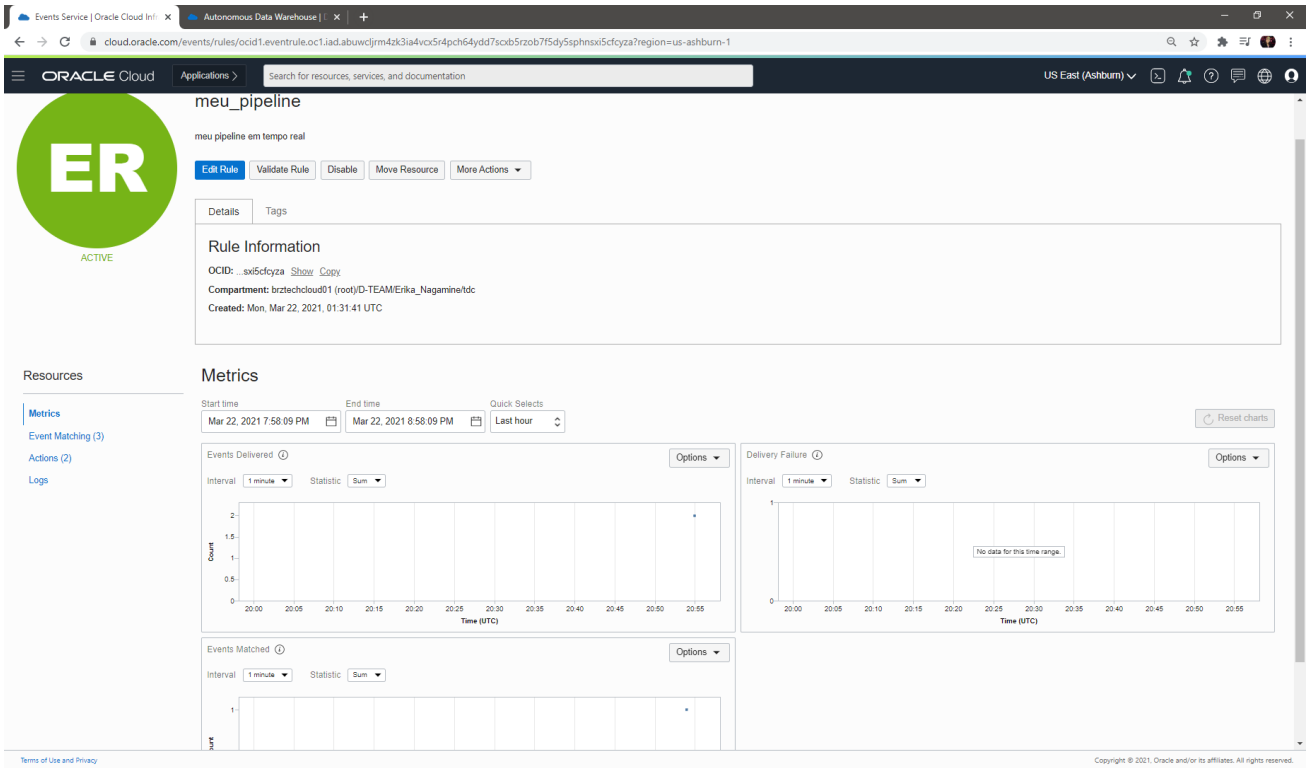
Ao final do processo você terá o arquivo dentro do bucket processed bucket:



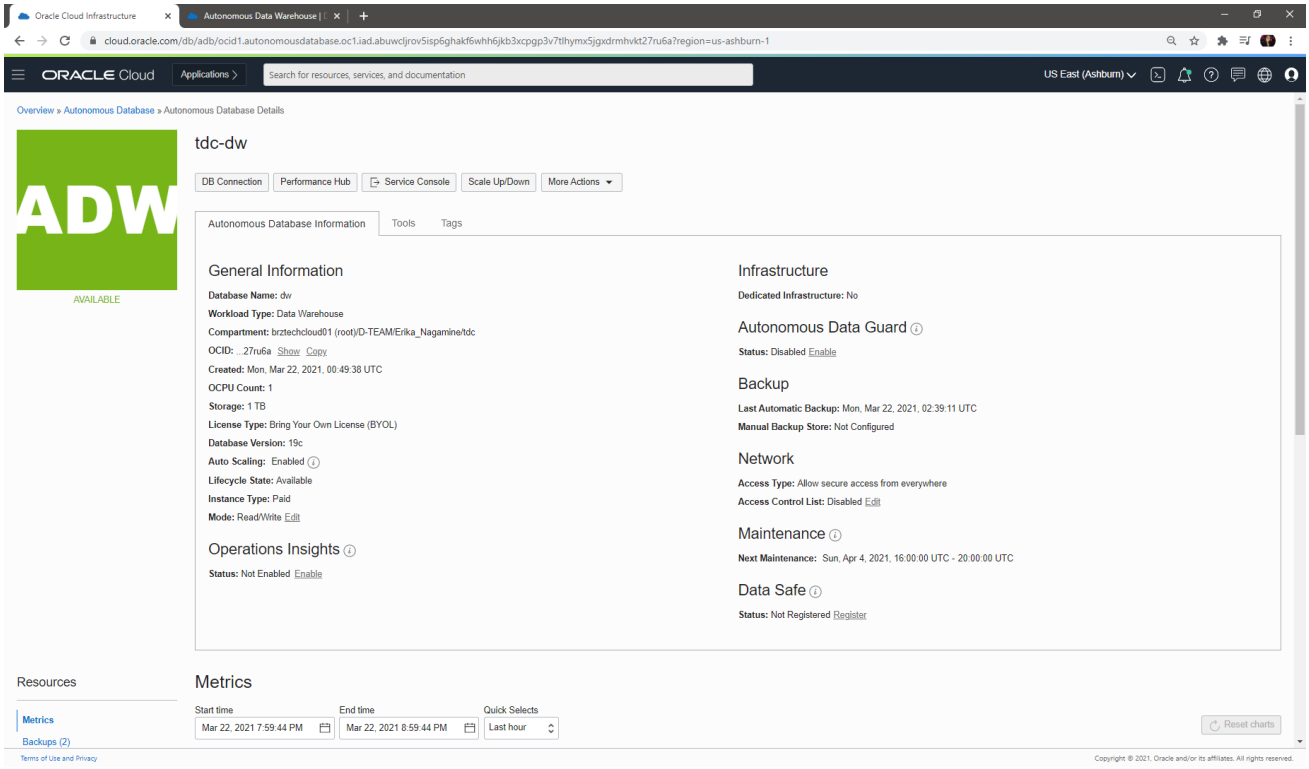
Note que seu processo estará nas métricas da função:



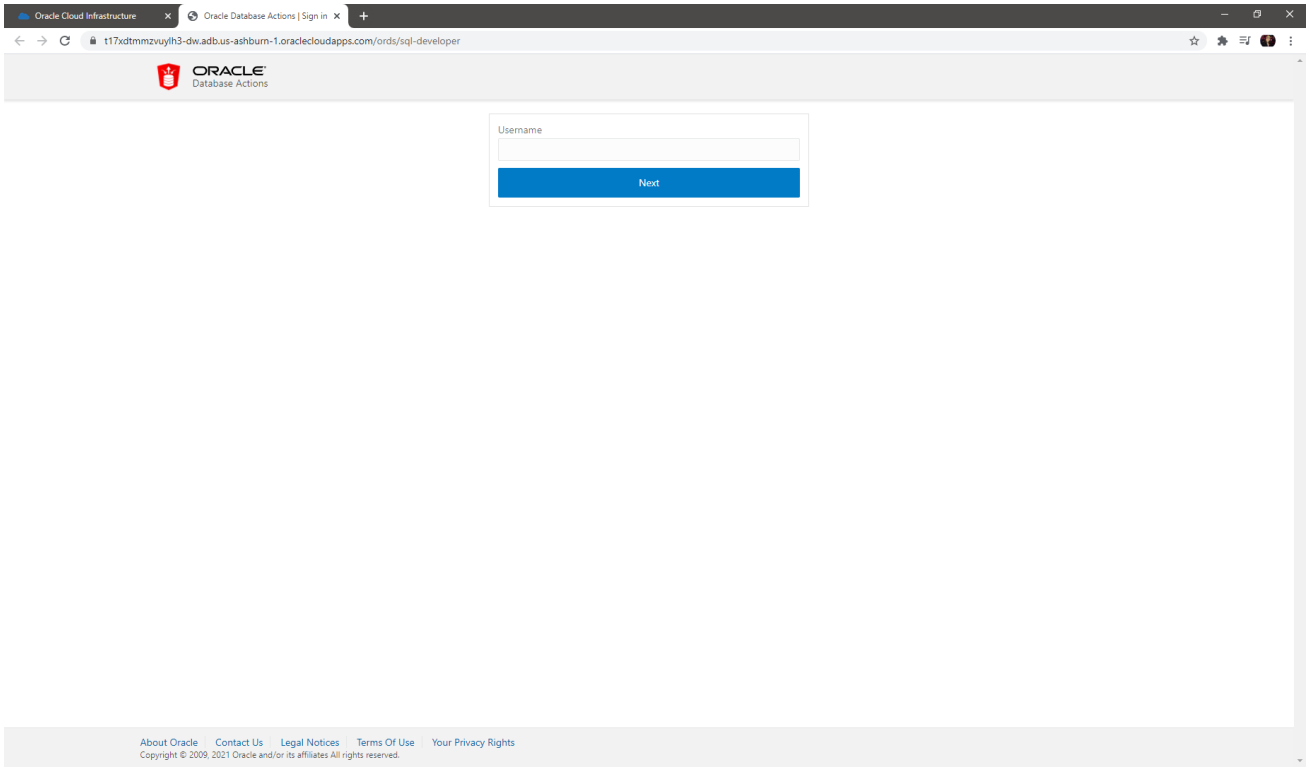
E nas métricas do evento:



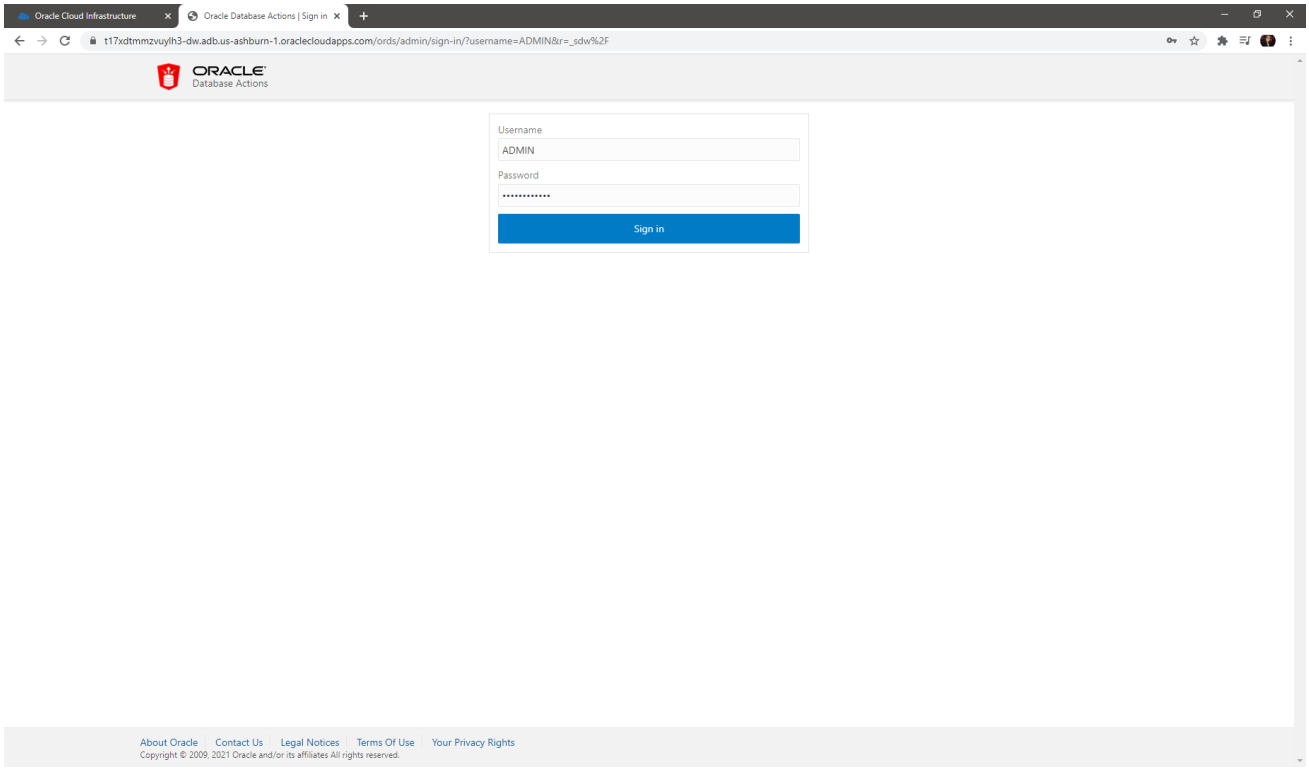
Você também pode fazer a validação dentro do seu dw, para isso vá no ADW:



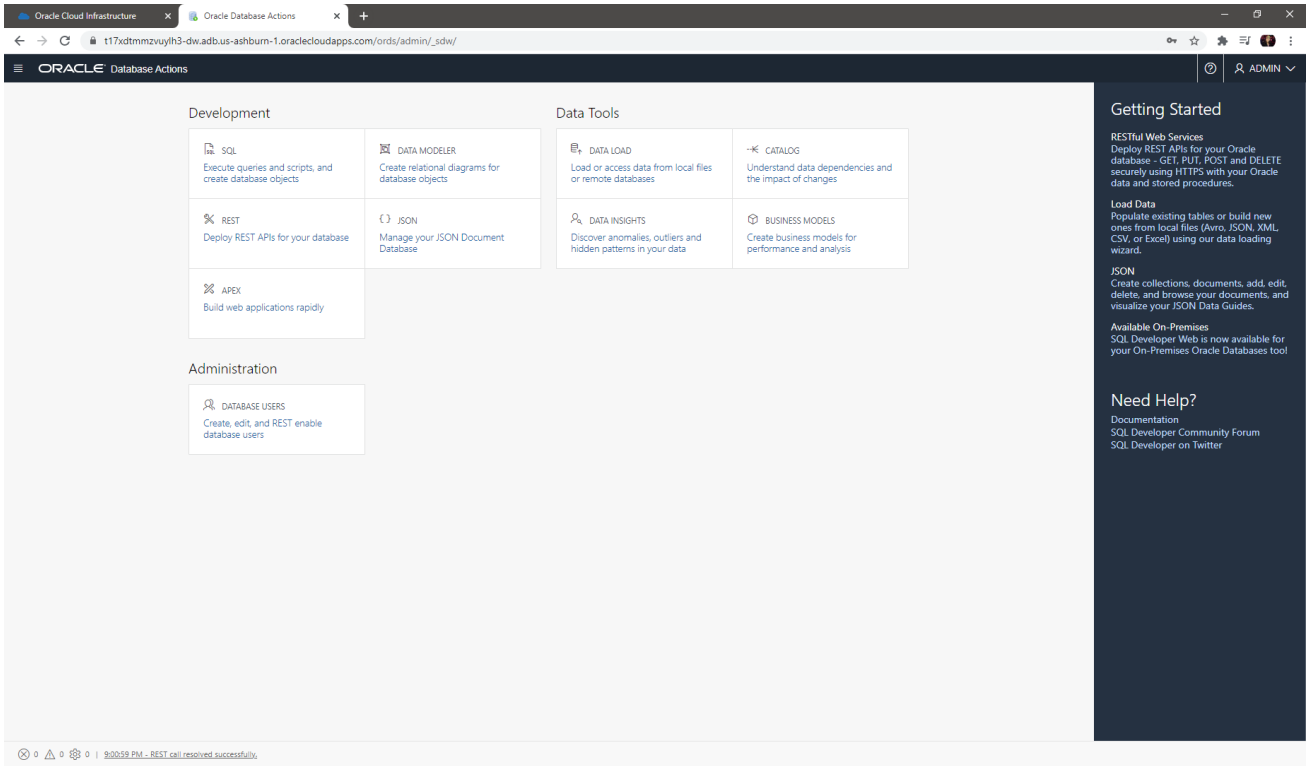
Clique em tools -> database actions:



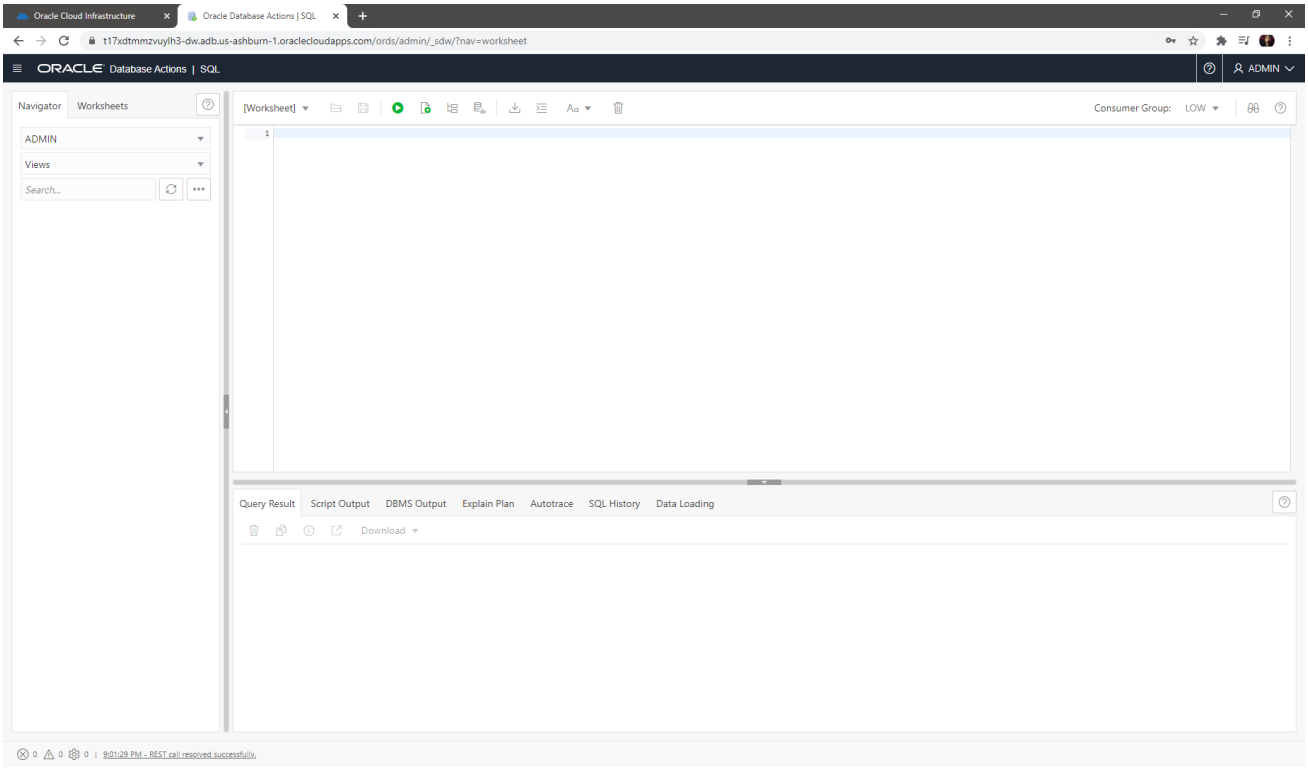
Insira seu usuário e senha:



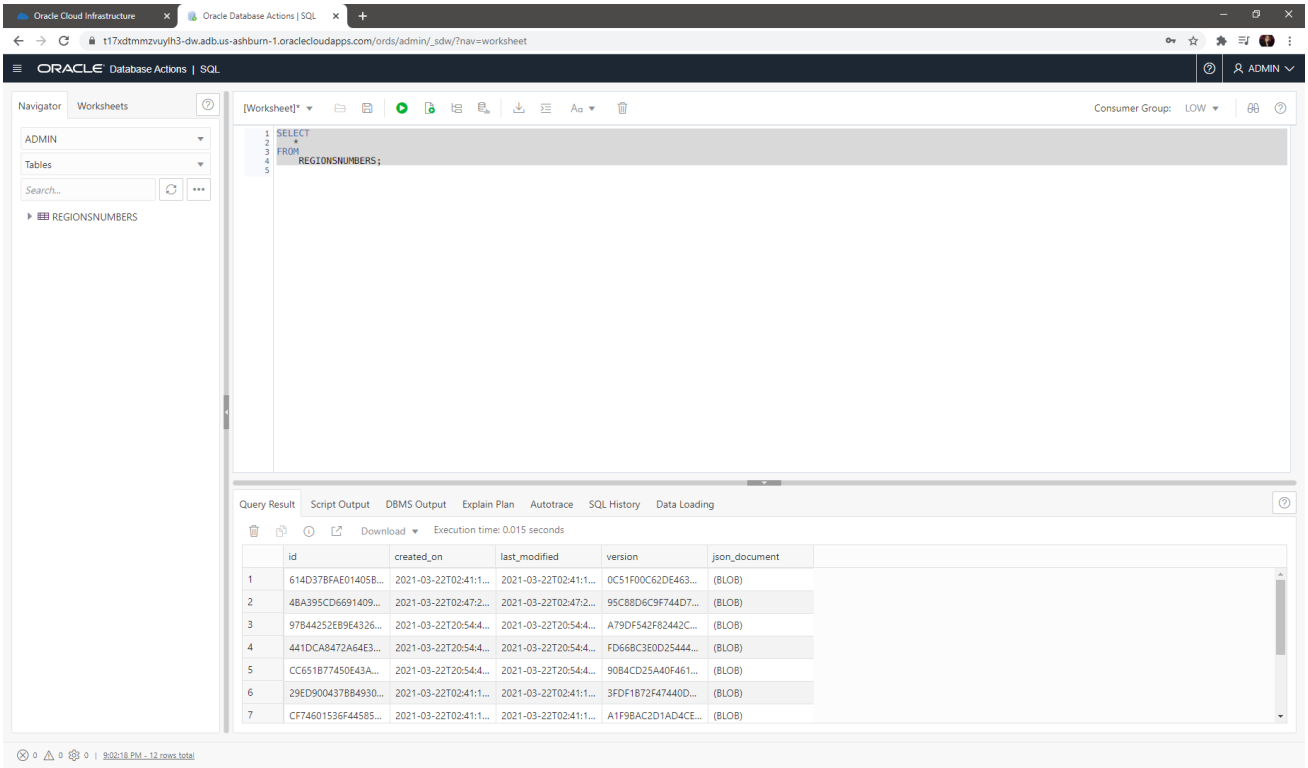
Clique em SQL:



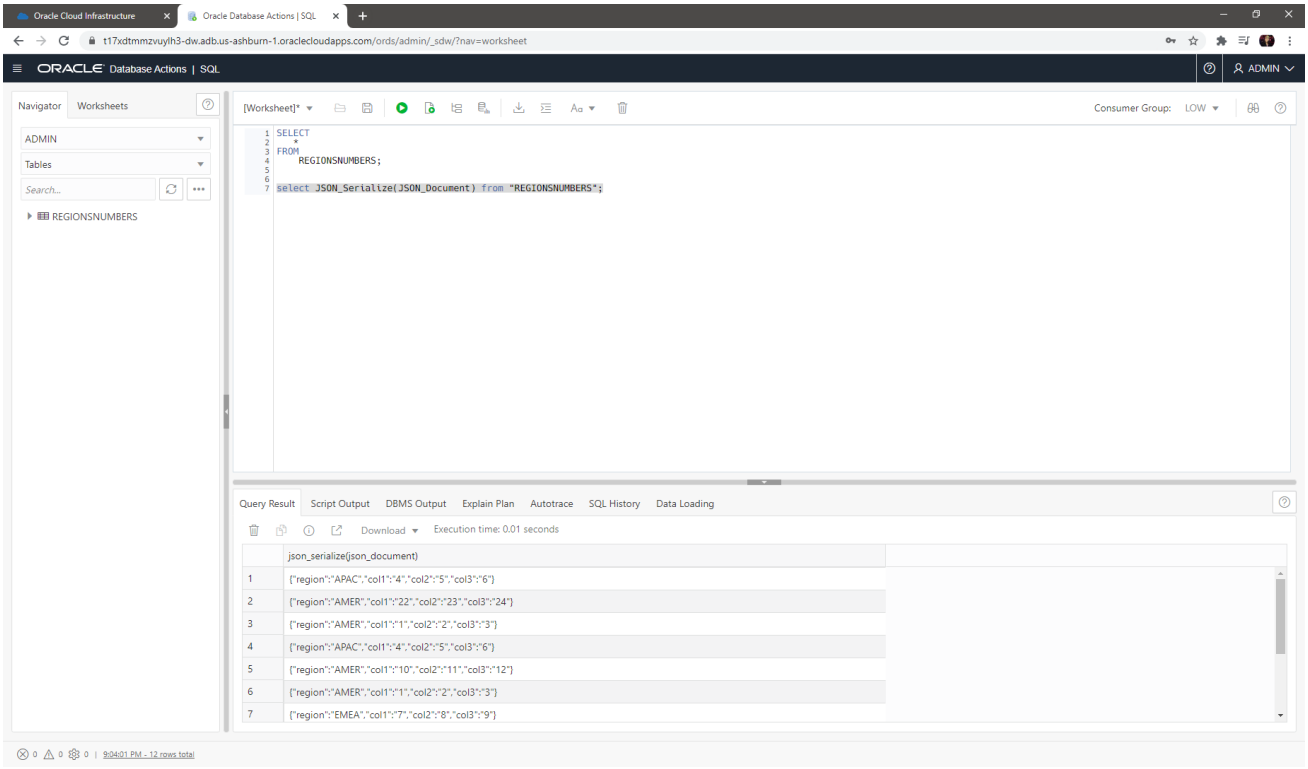
Na tela de SQL, faça a query no objeto criado:



Ao realizar q query note que ela aparece como objeto json:



Veja que é possível melhorar a visualização utilizando os recursos json:


Oracle Cloud Infrastructure
Oracle Database Actions | SQL
t17xdtmmzvuylh3-dw.adb.us-ashburn-1.oraclecloudapps.com/ords/admin/_sdw/?nav=worksheet
ORACLE Database Actions | SQL
ADMIN
Navigator
Worksheets
ADMIN
Tables
Search...
REGIONSNUMBERS
[Worksheet]*
Consumer Group: LOW
SELECT
*
FROM
REGIONSNUMBERS;
select JSON_Serialize(JSON_Document) from "REGIONSNUMBERS";
Query Result
Script Output
DBMS Output
Explain Plan
Autotrace
SQL History
Data Loading
Download
Execution time: 0.01 seconds
json_serialize(json_document)
1 {"region":"APAC","col1":"4","col2":"5","col3":"6"}
2 {"region":"AMER","col1":"22","col2":"23","col3":"24"}
3 {"region":"AMER","col1":"1","col2":"2","col3":"3"}
4 {"region":"APAC","col1":"4","col2":"5","col3":"6"}
5 {"region":"AMER","col1":"10","col2":"11","col3":"12"}
6 {"region":"AMER","col1":"1","col2":"2","col3":"3"}
7 {"region":"EMEA","col1":"7","col2":"8","col3":"9"}
9:04:01 PM - 12 rows total

5. Referências

Algumas documentações importantes:

- Functions: https://www.oracle.com/webfolder/technetwork/tutorials/infographics/oci_functions_cloudshell_quickview/functions_quickview_top/functions_quickview/index.html#
- OCI Examples: https://github.com/oracle/oracle-functions-samples/tree/master/samples/oci-load-file-into-adw-python
- Archicture Reference: https://docs.oracle.com/en/solutions/serverless-dataload-adw/index.html#GUID-36A94B9D-0B06-4A4A-81B0-A292392181BD