

# Lab 2.

## Ingestão e Transformação de Dados com OCI Data Flow



Este Laboratório apresenta como podemos criar e executar a sua primeira aplicação utilizando o Data Flow Application no OCI.

## OCI Data & AI Fast Track – Hands-on Lab

### Validação Pré Requisitos

Nesta etapa iremos validar se alguns pré-requisitos necessários para a criação de Data Flow applications foram criados corretamente e já estão disponíveis.

Entre os principais pré-requisitos do Data Flow, e também recursos necessários para a execução desta atividade podemos destacar:

- Buckets utilizados pelo Data Flow;
- Buckets de input e outputs de dados;
- Identificação do Namespace do object Storage;
- Download do script python de exemplo;
- Download do Dataset listings\_summary.csv e reviews\_summary.csv em formato csv.

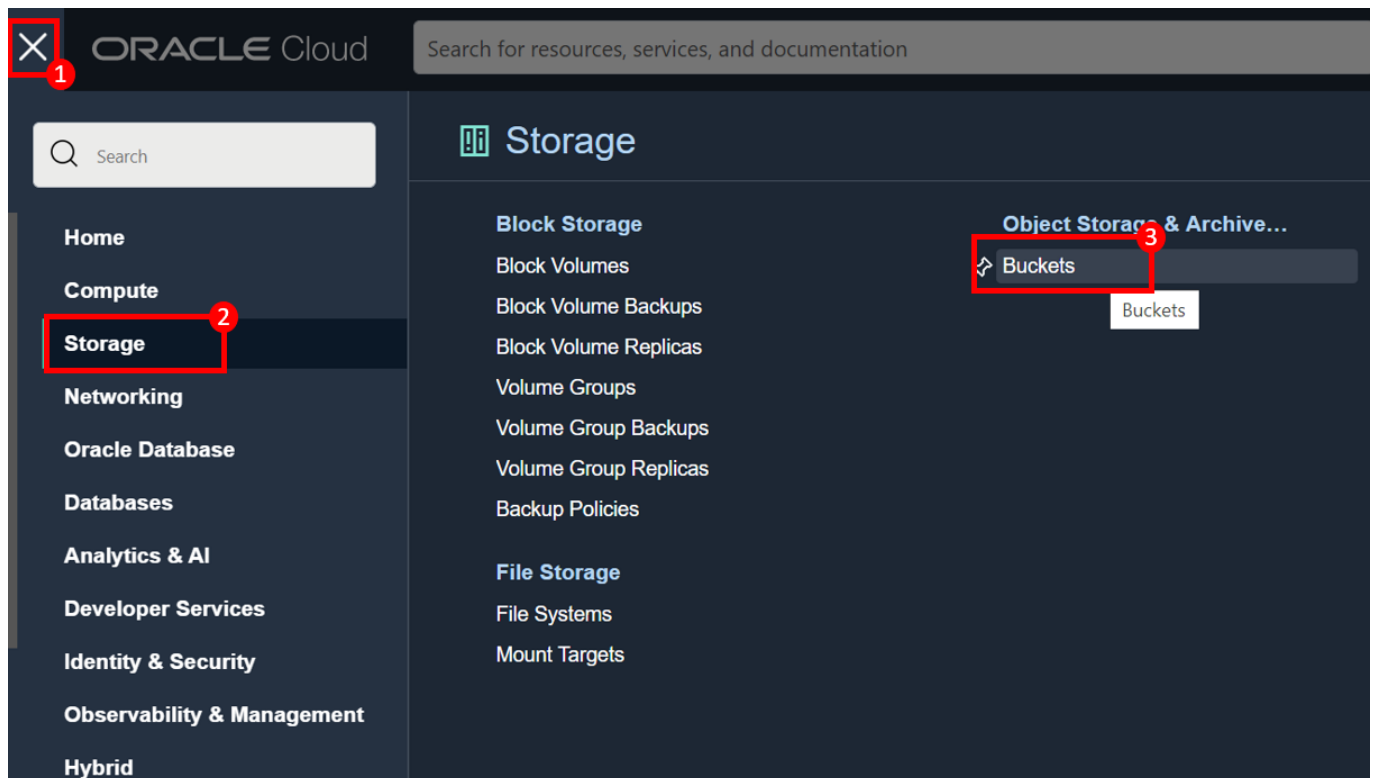
### Verificação dos buckets necessários

Nessa etapa, vamos validar a existência/criação dos seguintes buckets no object storage:

- dataflow-logs
- dataflow-app
- raw-data
- data-out

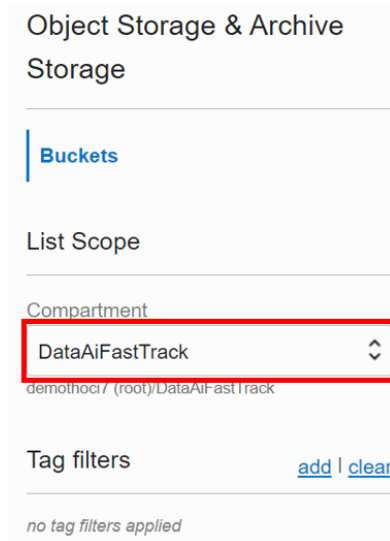
### Acessando o OCI Object Storage

Utilizando o menu de hambúguer, no canto superior esquerdo. Em seguida, selecione Storage e dentro de Object Storage & Archive clique em Buckets:



## OCI Data & AI Fast Track – Hands-on Lab

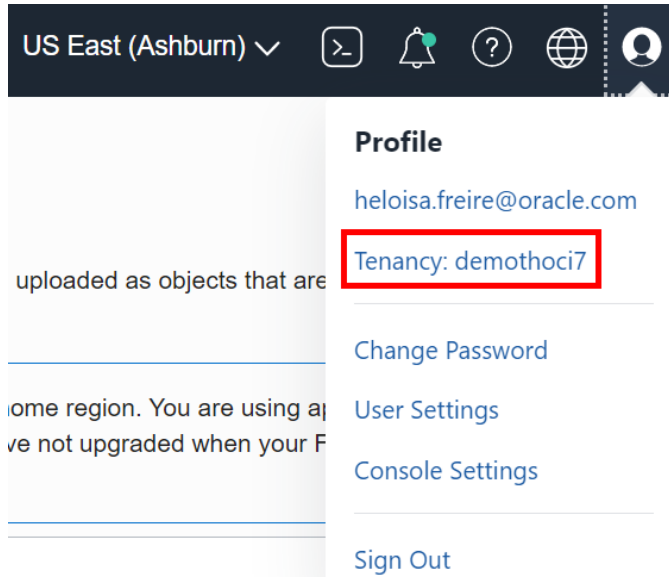
Verifique ao lado esquerdo da tela se estamos no Compartimento correto (DataAiFastTrack), devemos neste ponto visualizar todos os buckets listados acima:



### Identificação do Namespace do Object Storage

Nessa etapa, coletar o Namespace do Object Storage do seu ambiente. Esta informação é de extrema relevância, pois será utilizada nas etapas de configuração do nosso script python.

Para visualizar e copiar o Namespace do seu ambiente, acesse o menu com seu avatar de usuário no canto superior direito, e clique no nome do seu Tenancy:

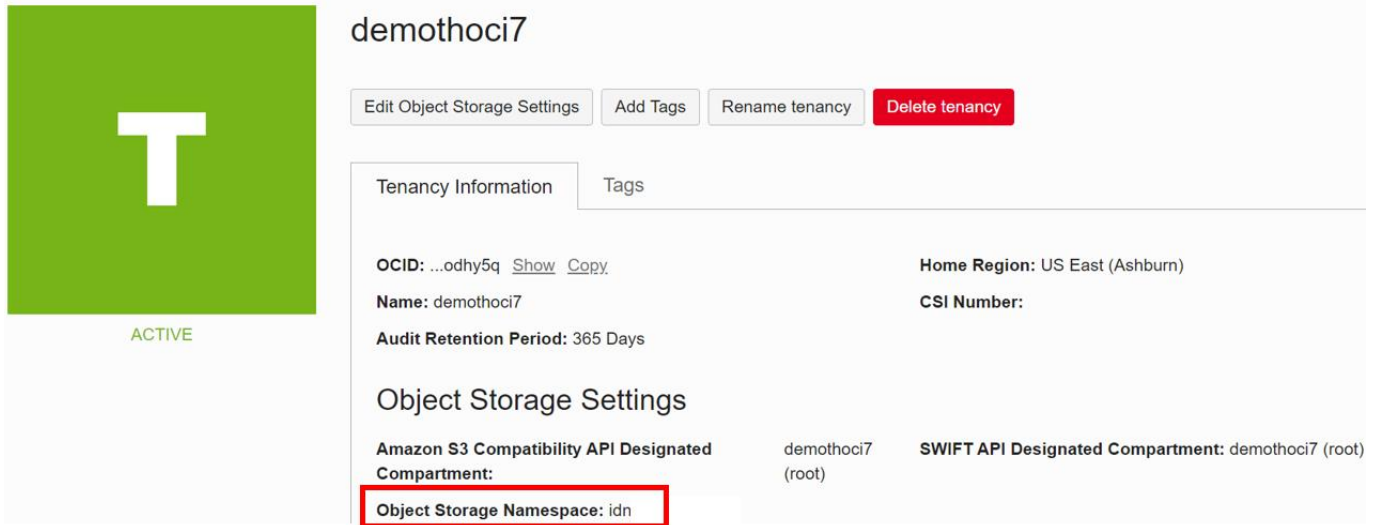


## OCI Data & AI Fast Track – Hands-on Lab

Agora nas informações do seu Tenancy, podemos encontrar e copiar do Object Storage Namespace.

*Atenção: Guarde o nome do Namespace em um notepad ou editor de sua preferência.*

Administration » Tenancy details



### Download dos Arquivos Utilizados

Durante esse LAB iremos utilizar dos arquivos, o script python **csv\_to\_parquet.py** e os datasets **listings\_summary.csv** e **reviews\_summary.csv**. Esses arquivos, você pode localizar no link abaixo:

<https://github.com/heloisaescobar/FastTrack-Data-AI>

### Configurando o Script Python

Após o download dos arquivos no passo anterior, primeiramente vamos trabalhar com o `csv_to_parquet.py`. Abra esse arquivo no editor de texto de sua preferência.

Dentro do Script, localize a variável `NAMESPACE`.

Agora utilizaremos o namespace já identificado para adicionar o valor nessa variável.

```
def main():  
  
    # Add your Namespace  
    NAMESPACE = "<++++NAMESPACE++++>"  
  
    # TODO: Set input and output paths.  
    INPUT_PATH1 = "oci://raw-data@" + NAMESPACE + "/reviews_summary.csv"  
    OUTPUT_PATH1 = "oci://data-out@" + NAMESPACE + "/reviews_summary.parquet"  
  
    INPUT_PATH2 = "oci://raw-data@" + NAMESPACE + "/listings_summary.csv"  
    OUTPUT_PATH2 = "oci://data-out@" + NAMESPACE + "/listings_summary.parquet"
```

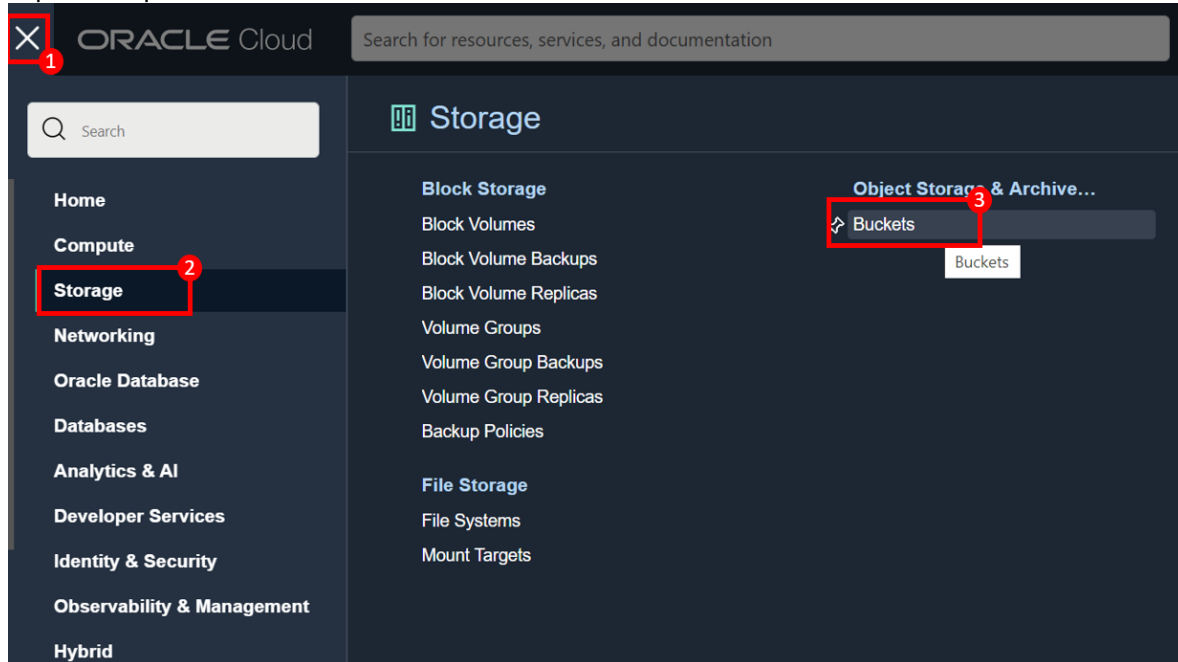
*Atenção: colocar o seu namespace entre "" (aspas duplas).*

## OCI Data & AI Fast Track – Hands-on Lab

### Transferindo arquivos utilizados para os buckets

Nessa etapa vamos utilizar a própria UI do OCI para fazer o upload dos arquivos para os buckets corretos no Object Storage.

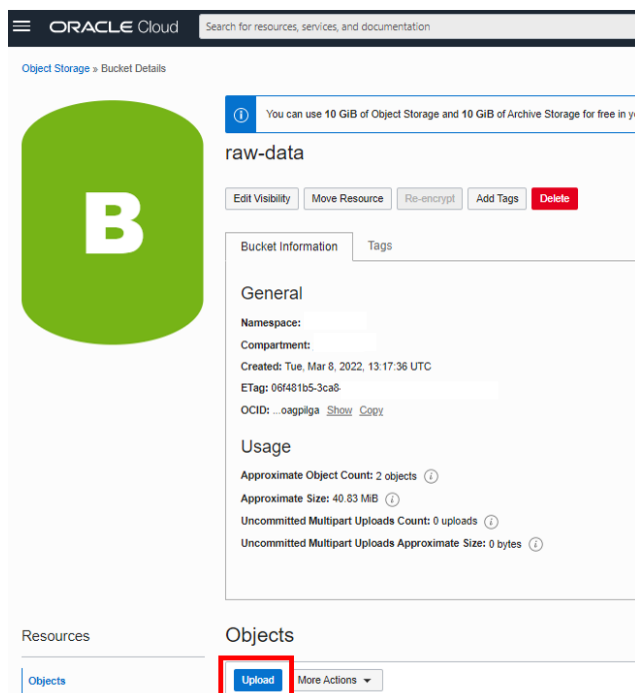
Para realizar esta transferência devemos acessar o serviço do Object Storage a partir do menu de hamburguer no canto superior esquerdo.



### Transferindo os Datasets: listings\_summary.csv e reviews\_summary.csv

Para que possamos fazer a transformação, iremos utilizar os datasets listings\_summary.csv e reviews\_summary.csv no buckets raw-data.

Para realizar o upload, acessar o buckets desejado (raw-data), e clicar no botão UPLOAD e transferir os arquivos:



## OCI Data & AI Fast Track – Hands-on Lab

### Upload Objects [Help](#)

Object Name Prefix *Optional*

Storage Tier

Standard

Choose Files from your Computer

Drop files here or [select files](#)


[Show Optional Response Headers and Metadata](#)

[Upload](#) [Cancel](#)

### Transferindo o Script Python para o Bucket

Seguindo o mesmo procedimento executado para o upload dos datasets, agora vamos transferir o script python com as alterações já realizadas csv\_to\_parquet.py para o bucket dataflow-app.

Object Storage > Bucket Details



**dataflow-app**

[Edit Visibility](#) [Move Resource](#) [Re-encrypt](#) [Add Tags](#) [Delete](#)

Bucket Information | Tags

**General**

Namespace:

Compartment:

Created: Tue, Mar 8, 2022, 13:17:43 UTC

ETag: 34e33c2a-fb58

OCID: ...u3ou57rq [Show](#) [Copy](#)

**Usage**

Approximate Object Count: 1 objects

Approximate Size: 3.13 KiB

Uncommitted Multipart Uploads Count: 0 uploads

Uncommitted Multipart Uploads Approximate Size: 0 bytes

Resources

Objects

[Upload](#) [More Actions](#)

## OCI Data & AI Fast Track – Hands-on Lab

### Upload Objects

[Help](#)

Object Name Prefix *Optional*

Storage Tier

Standard


Choose Files from your Computer

Drop files here or [select files](#)

[Show Optional Response Headers and Metadata](#)

Upload

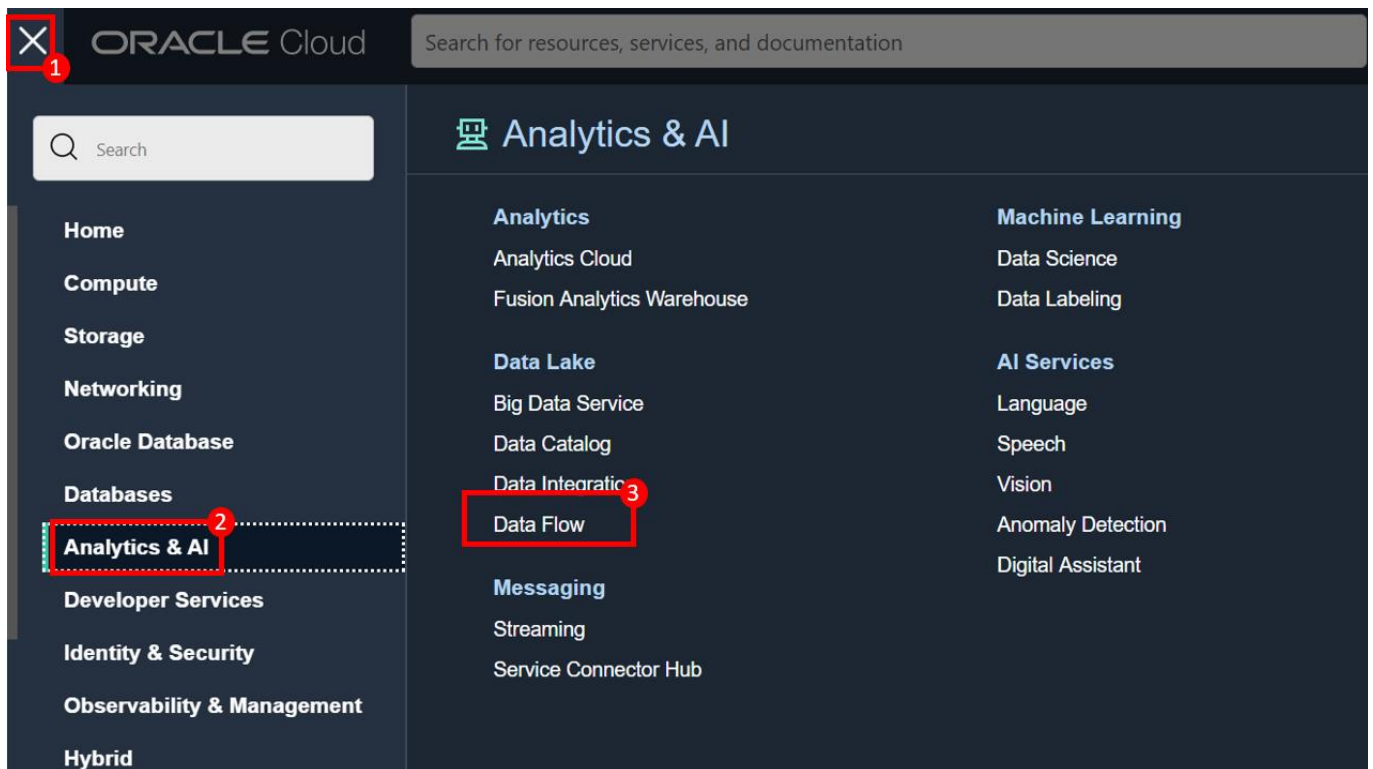
Cancel



### Criando sua Primeira Aplicação no OCI Data Flow

Neste passo iremos criar nossa primeira Aplicação no Data Flow. O objetivo desta aplicação será transformar os datasets listings\_summary.csv e reviews\_summary.csv em arquivos parquet, que poderá ser consumido por outras ferramentas.

Para acessarmos o DataFlow, seguidos ao menu de hamburguer no canto superior esquerdo e depois acessamos Analytics & IA e Data Flow.



## OCI Data & AI Fast Track – Hands-on Lab

Dentro da console do Data Flow, vamos clicar em *Create Application*.

Data Flow

[Applications](#)  
[Runs](#)  
[Private endpoints](#)

List Scope

Compartment  
DataAiFastTrack  
demothoci7 (root)/DataAiFastTrack

Tag Filters  
[add](#) | [clear](#)

### Applications *in* DataAiFastTrack *Compartment*

Data Flow lets you run Apache Spark jobs at any scale with almost no administration. [Learn more](#)

!

Oracle Data Flow prerequisites  
Before you can create, manage, and execute applications in Oracle Data Flow, your tenancy administrator should setup the following storage and policies:  
[Show more information](#)

Create application

Create sample application

Name	Language	Spark version	Language	Owner	Created	Updated
No items found.						

Showing 0 Items < Page 1 >

Após clicar no botão, iremos fornecer as informações necessárias para a criação da nossa aplicação.

## Create application

[Help](#)

### General information

Name ⓘ

csv2p

Description *Optional* ⓘ

Converter datasets csv para parquet

### Resource configuration

Spark version

Spark 3.0.2 -> Scala 2.12

Driver shape ⓘ

VM.Standard2.1 (15 GB Memory, 1 OCPU, 175 GB Block Volume)

Executor shape ⓘ

VM.Standard2.1 (15 GB Memory, 1 OCPU, 175 GB Block Volume)

Number of executors ⓘ

1



## OCI Data & AI Fast Track – Hands-on Lab

### Application configuration

☐ Spark streaming

☐ Use Spark-Submit Options [\(i\)](#)

Language

☐ Java ☒ Python ☐ SQL ☐ Scala

Select a file [\(i\)](#)

☐ Enter the file URL manually

Object Storage file name in **fast-track-4** [\(Change Compartment\)](#)

dataflow-app

csv\_to\_parquet\_teste.py

Create

[Cancel](#)

### Executando sua primeira Application – “Run”

Após criação de sua *Data Flow App*, agora podemos executar o código quantas vezes necessários através do botão RUN na console.

<div>Create application</div> <div>Create sample application</div>						
Name	Language	Spark version	Language	Owner	Created	Updated
<a href="#">csv2p</a>	Python	3.0.2	Batch	helois a.freire @oracl e.com	Tue, Mar 8, 2022, 14:05:25 UTC	Tue, Mar 8, 2022, 14:05:25 UTC
Showing 1 Item < Page 1 >						

## OCI Data & AI Fast Track – Hands-on Lab

Para cada execução podemos definir individualmente os parâmetros relacionados a infraestrutura alocada ou argumentos.

Para nossa execução, vamos clicar no botão *RUN*.

Data Flow > Applications > Application details

csv2p

[Edit](#) [Run](#) [Add Tags](#) [Move Resource](#) [Delete](#)

Application Information Tags

**Application configuration**

Language: Python  
Description: Converter para parquet [Show](#) [Copy](#)  
OCID: ...naulq [Show](#) [Copy](#)  
Application type: Batch  
Metastore: No value  
Default managed table location: No value  
File URL: ...to\_parquet\_teste.py [Show](#) [Copy](#)  
Archive URI: No value  
Arguments: No Value  
Application log location: ...w-logs@jdrnb64pt6/ [Show](#) [Copy](#)

**Resource configuration**

Spark version: 3.0.2  
Scala version: Scala 2.12  
Driver shape: VM.Standard2.1  
Executor shape: VM.Standard2.1  
Number of executors: 1

**Network**

Access type: Internet access (No subnet)

Vamos manter as configurações já configuradas na própria aplicação e clicar em *RUN*.

### Run Python application

Name ⓘ

csv2p

Driver shape ⓘ

VM.Standard2.1 (15 GB Memory, 1 OCPU, 175 GB Block Volume)

Executor shape ⓘ

VM.Standard2.1 (15 GB Memory, 1 OCPU, 175 GB Block Volume)

Number of executors ⓘ

1

Arguments Optional ⓘ

Tags

Tagging is a metadata system that allows you to organize and track resources within your tenancy. Tags are composed of keys and values that can be attached to resources.

[Learn more about tagging](#)

Tag Namespace	Tag Key	Value
None (add a free-form tag)		

+ Additional Tag

Show advanced options

[Run](#) [Cancel](#)

# OCI Data & AI Fast Track – Hands-on Lab

## Verificando Logs e o Resultado Esperado

O Dataflow registra automaticamente os logs de erro dos nodes driver e executors no bucket dataflow-logs, os logs de stderr e stdout são gerados a cada execução em dois arquivos .gz diferentes.

As logs ficam listadas a cada *RUN* das Aplicações, porém esta exibição não ocorre de imediato. Caso necessário verificar uma log que ainda não esteja sendo exibida na console de execução, podemos também verificar esta log diretamente no bucket dataflow-logs.

R

SUCCEEDED

csv2p on Tue, Mar 8, 2022, 14:14:12 UTC

Re-run

Spark UI

Add Tags

Move Resource

Stop

Run information

Tags

Run information

State details: Not available

Created: Tue, Mar 8, 2022, 14:14:12 UTC

Owner: heloise.freire@oracle.com

Request Id: ...4ff43365e Show Copy

Resource configuration

Spark version: 3.0.2

Scale: Scala 2.12

Driver shape: VM.Standard2.1

Executor shape: VM.Standard2.1

Number of executors: 1

Network

Access type: Internet access (No subnet)

Application configuration

Application Name: csv2p... Show Copy

Language: Python

Description: Converter para parquet

OCID: ...gaova Show Copy

File URL: ...to\_parquet\_teste.py Show Copy

Archive URI: No value

Application type: Batch

Metastore: No value

Default managed table location: No value

Arguments: No Value

Application log location: oci://dataflow-logs@idnb64p66/ Show Copy

Resources

Related runs

Metrics

Logs

Logs

Archive logs

Name	File size	Source	Type	Created
<a href="#">spark_application_stdout.log.gz</a>	20 bytes	APPLICATION	STDOUT	Tue, Mar 8, 2022, 14:17:00 UTC
<a href="#">spark_application_stderr.log.gz</a>	20 bytes	APPLICATION	STDERR	Tue, Mar 8, 2022, 14:17:00 UTC
<a href="#">spark_driver_stderr_20220308T141000Z.log.gz</a>	8 KB	DRIVER	STDERR	Tue, Mar 8, 2022, 14:29:05 UTC
<a href="#">spark_executor_stderr_20220308T141000Z.log.gz</a>	7 KB	EXECUTOR	STDERR	Tue, Mar 8, 2022, 14:30:03 UTC

Após a execução com sucesso, podemos então verificar o bucket data-out e confirmar a geração dos arquivos .parquet conforme esperado:

Edit Visibility

Move Resource

Re-encrypt

Add Tags

Delete

Bucket Information

Tags

General

Namespaces: idnb64p66

Compartment: [fast-track-4](#)

Created: Tue, Mar 8, 2022, 13:17:49 UTC

Etag: a2be8db-e651-4f30-83c5-e37f37ee076c

OCID: ...wyg62w4q Show Copy

Usage

Approximate Object Count: 0 objects

Approximate Size: 6.72 MB

Uncommitted Multipart Uploads Count: 0 uploads

Uncommitted Multipart Uploads Approximate Size: 0 bytes

Features

Default Storage Tier: Standard

Visibility: Private

Encryption Key: Oracle managed key [Assign](#)

Auto-Tiering: @ Disabled Edit ?

Emit Object Events: @ Disabled Edit ?

Object Versioning: @ Disabled Edit ?

Objects

Upload

More Actions

Name

Last Modified

Size

Storage Tier

<div><div></div><div>insights_summary.parquet</div></div>	-	-	-
<div><div></div><div>._SUCCESS</div></div>	Tue, Mar 8, 2022, 14:17:05 UTC	0 bytes	Standard
<div><div></div><div>part-00000-fdb1d800-b046-4f26-96ba-67ceffa43a32-c000.snappy.parquet</div></div>	Tue, Mar 8, 2022, 14:17:04 UTC	1.95 MiB	Standard
<div><div></div><div>part-00001-fdb1d800-b046-4f26-96ba-67ceffa43a32-c000.snappy.parquet</div></div>	Tue, Mar 8, 2022, 14:17:05 UTC	1.41 MiB	Standard
<div><div></div><div>reviews_summary.parquet</div></div>	-	-	-
<div><div></div><div>._SUCCESS</div></div>	Tue, Mar 8, 2022, 14:16:57 UTC	0 bytes	Standard
<div><div></div><div>part-00000-0145489a-f19d-4dd5-88a2-c2c806a40437-c000.snappy.parquet</div></div>	Tue, Mar 8, 2022, 14:16:56 UTC	1.95 MiB	Standard
<div><div></div><div>part-00001-0145489a-f19d-4dd5-88a2-c2c806a40437-c000.snappy.parquet</div></div>	Tue, Mar 8, 2022, 14:16:56 UTC	1.41 MiB	Standard