



Titulación de Sistemas Informáticos y Computación

Implementación de un ambiente de Ciencia de Datos a través de la Librería Open Source Rapids

José Alberto Guarnizo Romero

Tutor: Mgtr. René Rolando Elizalde Solano

2020

Agenda

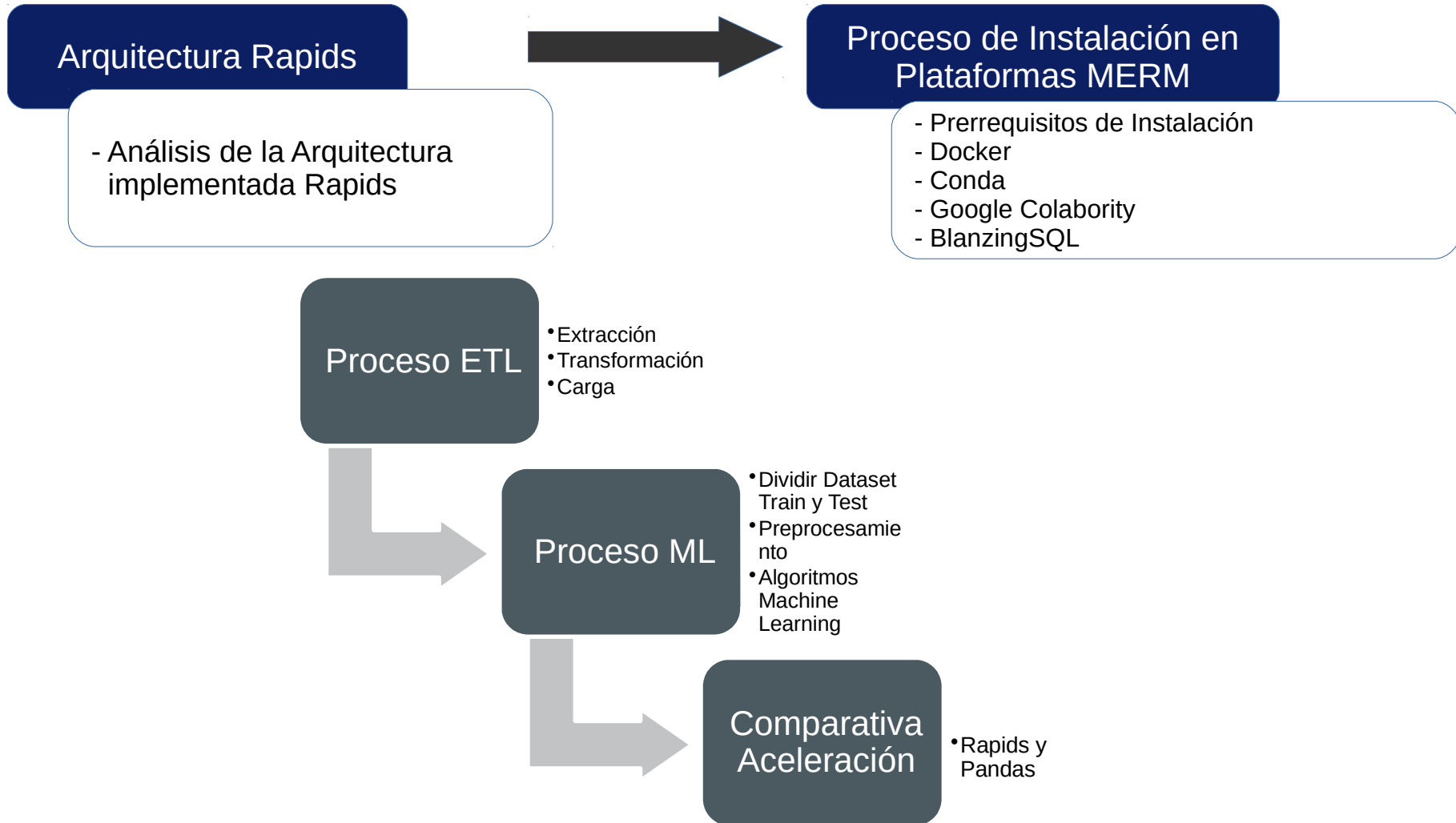
- Objetivos
- Arquitectura Rapids
- Plataformas MERM
- Proceso ETL
- Proceso ML
- Comparativa Aceleración

Objetivos del TT

- Objetivo General
 - Implementar un ambiente de Ciencia de Datos a través de la Librería Open Source Rapids.
- Objetivos Específicos
 - Elaborar una investigación documentada sobre la Librería Open Source Rapids.
 - Realizar analítica de grandes volúmenes de datos a través del uso de las características de la Librería Open Source Rapids.
 - Implementar algoritmos de Machine Learning haciendo uso de la Librería Open Source Rapids.

FASES DEL TRABAJO DE TITULACIÓN

Fases del Proyecto

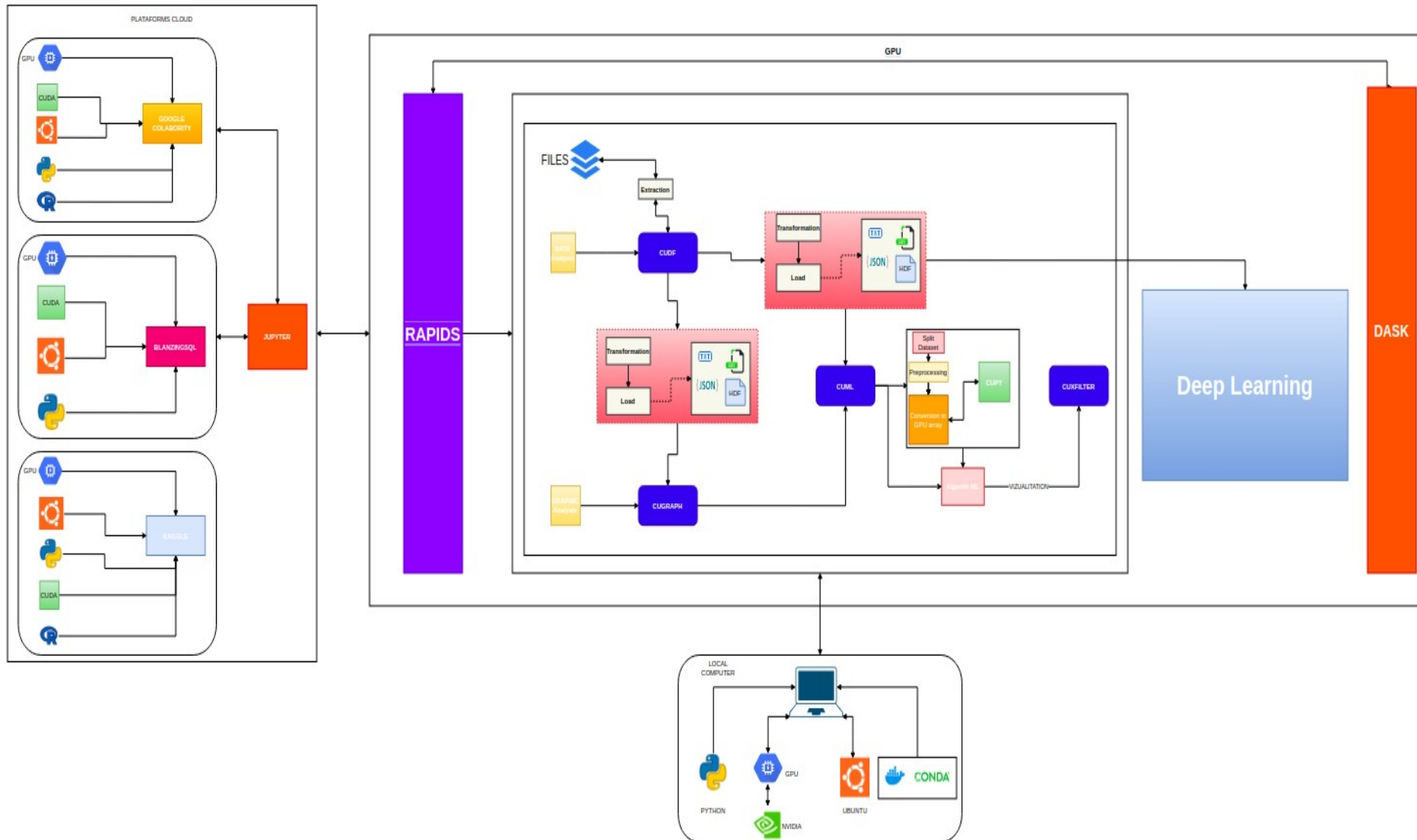


¿De verdad es necesaria la aceleración en GPU?



viernes 17 de julio de 2020

Arquitectura Rapids: Análisis de la Arquitectura implementada Rapids.



Proceso de Instalación en Plataformas

MERM: Prerrequisitos de Instalación

Prerrequisitos de Instalación localmente.

Gpu Nvidia	Sistema Operativo Linux (Ubuntu)	Docker	Nvidia Drivers Cuda
Titan RTX Tesla GeForce	Versión 16.04 Versión 18.04	Versión 19.03	Versión 10.0 Versión 10.1 Versión 10.2

Computador portátil local

Marca Portátil	Intel	Sistema Operativo	Gpu
Dell	Core I7 8th Gen	Ubuntu 18.04	Nvidia GeForce MX150 4GB

Proceso de Instalación en Plataformas

MERM: Docker.

```
root@jose-Inspiron-7472:/home/jose# docker run --gpus all nvidia/cuda:10.0-base nvidia-smi
Unable to find image 'nvidia/cuda:10.0-base' locally
10.0-base: Pulling from nvidia/cuda
7ddb4c47eeb70: Pull complete
c1bbdc448b72: Pull complete
8c3b70e39044: Pull complete
45d437916d57: Pull complete
d8f1569ddae6: Pull complete
de5a2c57c41d: Pull complete
ea6f04a00543: Pull complete
Digest: sha256:e6e1001f286d084f8a3aea991afbcfe92cd389ad1f4883491d43631f152f175e
Status: Downloaded newer image for nvidia/cuda:10.0-base
Tue Jul 7 05:25:39 2020

+-----+
| NVIDIA-SMI 440.100      Driver Version: 440.100      CUDA Version: 10.2      |
+-----+-----+
| GPU Name      Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|  Memory-Usage | GPU-Util  Compute M. |
+-----+-----+
| 0  GeForce MX150      Off | 00000000:01:00.0 Off |                 N/A |
| N/A   50C    P0      N/A   /   N/A     | 254MiB / 4042MiB |      0%    Default  |
+-----+-----+

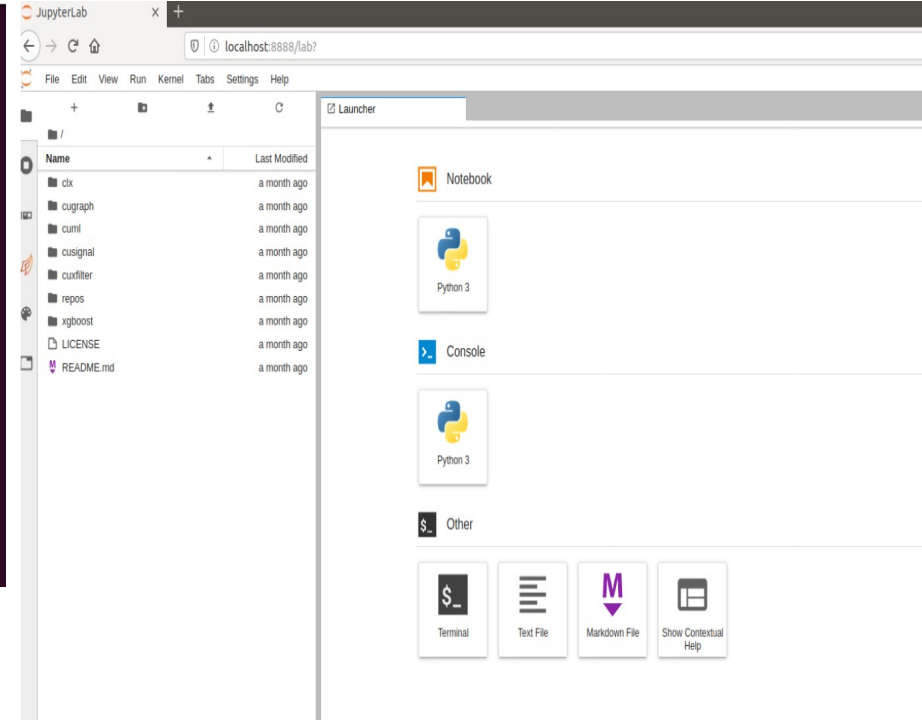
+-----+
| Processes:                        GPU Memory |
|   GPU       PID    Type    Process name                     Usage |
+-----+-----+

```

Nvidia de computador local
se encuentre en la imagen
docker

```
root@jose-Inspiron-7472:/home/jose# docker run --gpus all --rm -it -p 8888:8888 -p 8787:8787 -p 8786:8786 \
    rapidsai/rapidsai:cuda10.2-runtime-ubuntu18.04-py3.7
rapids) root@697a8b7466ee:/rapids/notebooks#
rapids) root@697a8b7466ee:/rapids/notebooks#
```

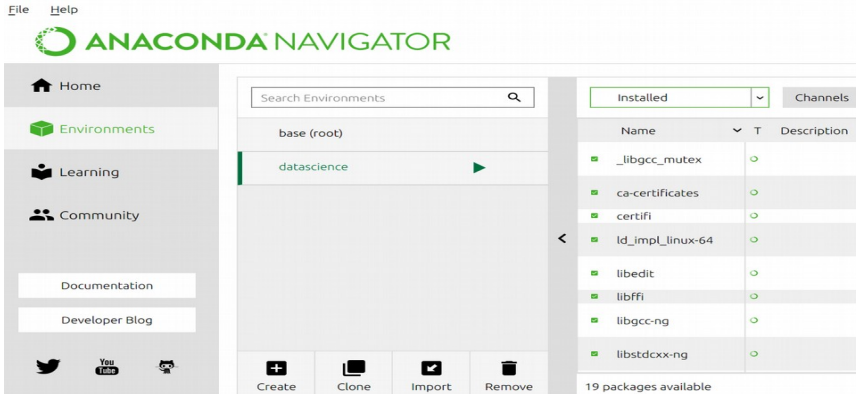
Ejecutando Imagen
docker



Plataforma de
Desarrollo

Proceso de Instalación en Plataformas

MERM: Conda.



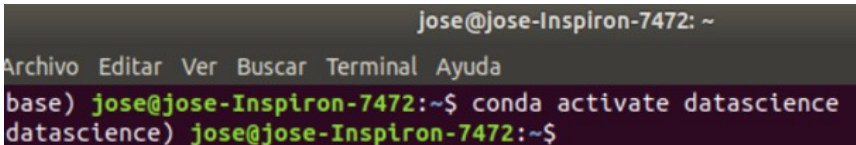
```

data science) jose@jose-Inspiron-7472:~$ jupyter notebook
I 23:10:03.116 NotebookApp] Writing notebook server cookie secret to /home/jose/.local/share/jupyter/runtime/notebook_cookie_secret
I 23:10:03.576 NotebookApp] Serving notebooks from local directory: /home/jose
I 23:10:03.576 NotebookApp] The Jupyter Notebook is running at:
I 23:10:03.576 NotebookApp] http://localhost:8888/?token=1d26d63dfeffe1f426a642344b1918a6619fa0cbd164f49d
I 23:10:03.576 NotebookApp] or http://127.0.0.1:8888/?token=1d26d63dfeffe1f426a642344b1918a6619fa0cbd164f49d
I 23:10:03.576 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
C 23:10:03.618 NotebookApp]

To access the notebook, open this file in a browser:
file:///home/jose/.local/share/jupyter/runtime/nbserver-24901-open.html
Or copy and paste one of these URLs:
http://localhost:8888/?token=1d26d63dfeffe1f426a642344b1918a6619fa0cbd164f49d
or http://127.0.0.1:8888/?token=1d26d63dfeffe1f426a642344b1918a6619fa0cbd164f49d

```

Ejecutar jupyter



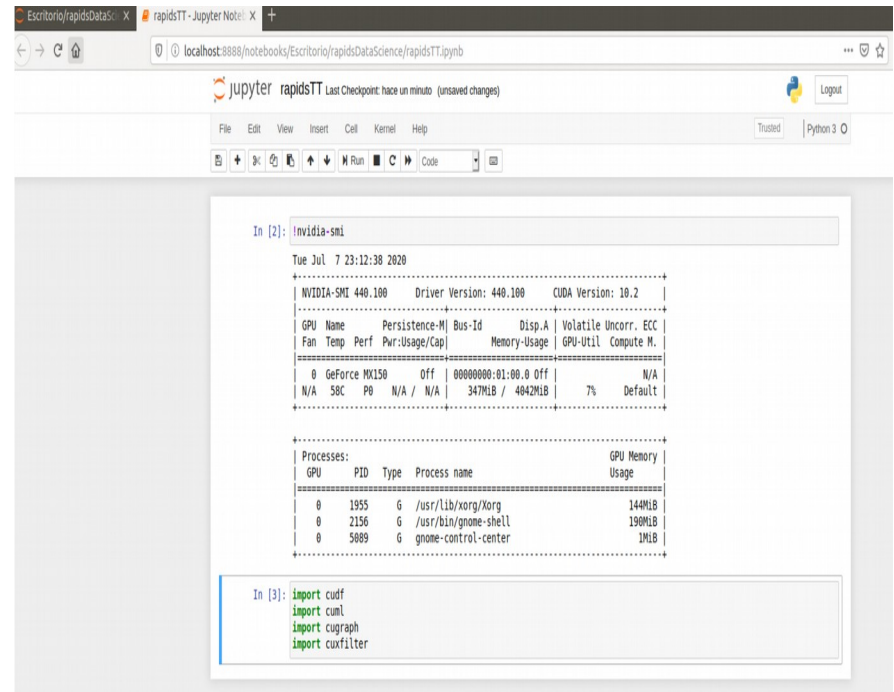
Activar entorno

```
(datascience) jose@jose-Inspiron-7472:~$ nvidia-smi
Tue Jul  7 22:09:06 2020

+-----+
| NVIDIA-SMI 440.100      Driver Version: 440.100      CUDA Version: 10.2      |
+-----+-----+
| GPU   Name               Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
+-----+-----+
|  0  GeForce MX150      Off          | 00000000:01:00:00 Off |          N/A         |
| N/A   60C    P0      N/A /  N/A     | 283MiB / 4042MiB |      5%      Default  |
+-----+-----+

+-----+
| Processes:                                                       GPU Memory |
|  GPU       PID    Type   Process name                     Usage    |
+-----+-----+
|    0      1955     G   /usr/lib/xorg/Xorg                135MiB   |
|    0      2156     G   /usr/bin/gnome-shell              145MiB   |
|    0      5089     G   gnome-control-center               1MiB    |
+-----+-----+
```

Verificar Nvidia



Plataforma de desarrollo

Proceso de Instalación en Plataformas

MERM: Google Colabority.

The screenshot displays the RAPIDS IDE interface. At the top, the 'RAPIDS' logo is visible. Below it, a menu bar includes 'Archivo', 'Editar', 'Ver', 'Insertar', 'Entorno de ejecución', 'Herramientas', 'Ayuda', and a link 'Se guardaron todos los cambios'. The main workspace is divided into two panes. The left pane shows a file explorer with a folder icon and a file named 'Código'. The right pane shows a code editor with a Python script. The script starts with a comment '[2] %tensorflow_version 2.' followed by an import statement 'import tensorflow as tf'. It then defines 'device_name = tf.test' and a conditional check 'if device_name != \"/dev/nvidia0\"'. If the condition is met, it raises a 'SystemError' with the message 'Found GPU at: /dev/nvidia0'. The script ends with a print statement 'print('Found GPU at: /dev/nvidia0')'. Below the code editor, there is a status bar showing 'Found GPU at: /dev/nvidia0'. A dropdown menu is open, showing various execution options and their corresponding keyboard shortcuts. The options are: 'Ejecutar todo' (Ctrl+F9), 'Ejecutar celdas anteriores a la seleccionada' (Ctrl+F8), 'Ejecutar la celda enfocada' (Ctrl+Enter), 'Ejecutar selección' (Ctrl+Shift+Enter), 'Ejecutar celda seleccionada y siguientes' (Ctrl+F10), 'Interrumpir la ejecución' (Ctrl+M), 'Reiniciar entorno de ejecución' (Ctrl+M), 'Reiniciar y ejecutar todo', 'Restablecer la configuración de fábrica del entorno de ejecución', and 'Cambiar tipo de entorno de ejecución'.

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda [Se guardaron todos los cambios](#)

+ Código + Texto

[2] %tensorflow_version 2.
import tensorflow as tf
device_name = tf.test
if device_name != '/dev/nvidia0':
 raise SystemError('Found GPU at: /dev/nvidia0')
print('Found GPU at: /dev/nvidia0')

Found GPU at: /dev/nvidia0

[3] !nvidia-smi

Thu Apr 30 15:41:41 2020

Ejecutar todo Ctrl+F9

Ejecutar celdas anteriores a la seleccionada Ctrl+F8

Ejecutar la celda enfocada Ctrl+Enter

Ejecutar selección Ctrl+Shift+Enter

Ejecutar celda seleccionada y siguientes Ctrl+F10

Interrumpir la ejecución Ctrl+M

Reiniciar entorno de ejecución Ctrl+M

Reiniciar y ejecutar todo

Restablecer la configuración de fábrica del entorno de ejecución

Cambiar tipo de entorno de ejecución

Configuración del notebook

Acelerador de hardware

GPU

Para aprovechar Colab al máximo, usa una GPU o incluso una TPU.

[Más información](#)

☐ Omitir el resultado

None

GPU

TPU

CANCELAR GUARDAR

```
# Install RAPIDS in colab
!git clone https://github.com/rapidsai/rapidsai-csp-utils.git
!bash rapidsai-csp-utils/colab/rapids-colab.sh stable

import sys, os

dist_package_index = sys.path.index('/usr/local/lib/python3.6/dist-packages')
sys.path = sys.path[:dist_package_index] + ['/usr/local/lib/python3.6/site-packages'] + sys.path[dist_package_index:]
sys.path
exec(open('rapidsai-csp-utils/colab/update_modules.py').read(), globals())
```

```
+ Código + Texto Conectar Editen
```

```
> |verification of version work  
|lsb_release -a  
|nvidia-smi
```

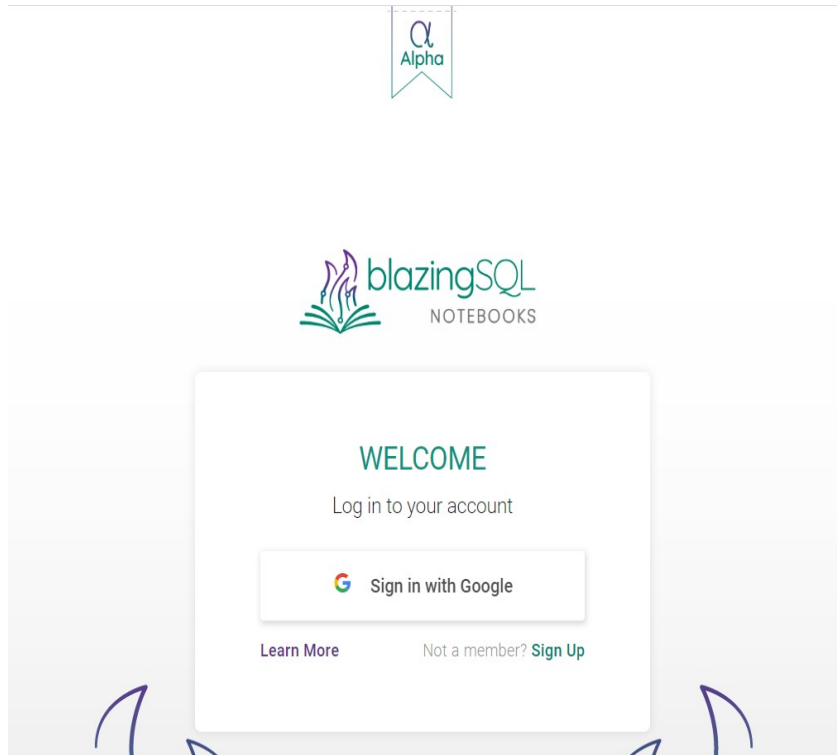
```
No LSB modules are available.  
Distributor ID: Ubuntu  
Description:    Ubuntu 18.04.3 LTS  
Release:        18.04  
Codename:       bionic  
Mon Jul 6 23:18:27 2020
```

NVIDIA-SMI 450.36.06 Driver Version: 418.67 CUDA Version: 10.1									
GPU Name		Persistence-M	Bus-Id	Disp.A	Volatile Uncomm. ECC				
Fan Temp Perf	PwrUsage/Cap	Memory-Usage	GPU-Util	Compute M.	MIG M.				
0 Tesla T4	Off	00000000:00:04:0 Off		0					
N/A 37C P8	9W / 70W	0MiB / 15079MiB	0%	Default	ERR!				

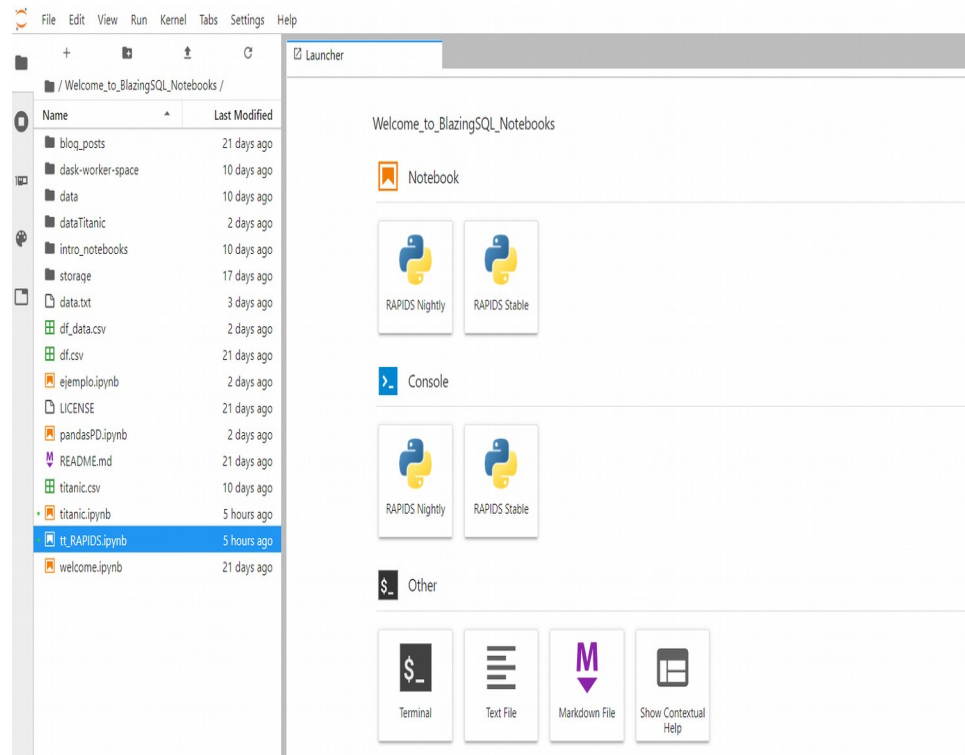
```
Processes:  
GPU GI CI PID Type Process name GPU Memory  
ID ID Usage
```

```
No running processes found
```

Proceso de Instalación en Plataformas MERM: BlazingSQL.



Página
BlazingSQL



Plataforma de
desarrollo

Dirección: <https://app.blazingsql.com/>

Proceso ETL: Extracción.

- Data con datos Semiestructurados: 80 mil filas de información a analizar de OpenCampus.

```
/edx/var/log/tracking/tracking.log-20171110-1510327021.gz:{"username": "", "event_source": "server", "name": "edx.course.enrollment.activated", "accept_langu
/edx/var/log/tracking/tracking.log-20171110-1510327021.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "66.249
/edx/var/log/tracking/tracking.log-20171110-1510330621.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "190.11
/edx/var/log/tracking/tracking.log-20171110-1510330621.gz:{"username": "", "event_type": "/i18n.js", "ip": "190.110.206.116", "agent": "Mozilla/5.0 (Windows NT
/edx/var/log/tracking/tracking.log-20171110-1510334221.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "172.17
/edx/var/log/tracking/tracking.log-20171110-1510334221.gz:{"username": "", "event_type": "/i18n.js", "ip": "172.17.24.79", "agent": "Mozilla/5.0 (Windows NT 6.3
/edx/var/log/tracking/tracking.log-20171110-1510341421.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "190.21
/edx/var/log/tracking/tracking.log-20171110-1510341421.gz:{"username": "", "event_type": "/i18n.js", "ip": "190.214.72.140", "agent": "Mozilla/5.0 (Windows NT
/edx/var/log/tracking/tracking.log-20171110-1510341421.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "181.11
/edx/var/log/tracking/tracking.log-20171110-1510341421.gz:{"username": "", "event_type": "/i18n.js", "ip": "181.113.136.34", "agent": "Mozilla/5.0 (Windows NT
/edx/var/log/tracking/tracking.log-20171110-1510341421.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "190.15
/edx/var/log/tracking/tracking.log-20171110-1510341421.gz:{"username": "", "event_type": "/i18n.js", "ip": "190.152.18.66", "agent": "Mozilla/5.0 (X11; Ubuntu; L
/edx/var/log/tracking/tracking.log-20171110-1510345021.gz:{"username": "FrankVac", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip"
/edx/var/log/tracking/tracking.log-20171110-1510345021.gz:{"username": "FrankVac", "event_type": "/i18n.js", "ip": "200.105.237.152", "agent": "Mozilla/5.0 (Wii
/edx/var/log/tracking/tracking.log-20171110-1510345021.gz:{"username": "FrankVac", "event_type": "/courses", "ip": "200.105.237.152", "agent": "Mozilla/5.0 (W
/edx/var/log/tracking/tracking.log-20171110-1510348621.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "40.77.1
/edx/var/log/tracking/tracking.log-20171110-1510352221.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "181.11
/edx/var/log/tracking/tracking.log-20171110-1510352221.gz:{"username": "", "event_type": "/i18n.js", "ip": "181.113.157.145", "agent": "Mozilla/5.0 (Windows NT
/edx/var/log/tracking/tracking.log-20171110-1510352221.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "181.11
/edx/var/log/tracking/tracking.log-20171110-1510352221.gz:{"username": "", "event_type": "/i18n.js", "ip": "181.113.157.145", "agent": "Mozilla/5.0 (Windows NT
/edx/var/log/tracking/tracking.log-20171110-1510352221.gz:{"username": "", "event_type": "/register", "ip": "181.113.157.145", "agent": "Mozilla/5.0 (Windows N
/edx/var/log/tracking/tracking.log-20171110-1510352221.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "172.16
/edx/var/log/tracking/tracking.log-20171110-1510352221.gz:{"username": "", "event_type": "/i18n.js", "ip": "172.16.44.31", "agent": "Mozilla/5.0 (Windows NT 6.1
/edx/var/log/tracking/tracking.log-20171110-1510366621.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "181.19
/edx/var/log/tracking/tracking.log-20171110-1510366621.gz:{"username": "", "event_type": "/i18n.js", "ip": "181.196.52.93", "agent": "Mozilla/5.0 (Windows NT 6
/edx/var/log/tracking/tracking.log-20171110-1510370221.gz:{"username": "", "event_type": "/courses/course-v1:UTPL+EFHE-Ed7+2017_DIC/about", "ip": "190.96
```

Data a Analizar

Proceso ETL: Extracción y lectura

RAPIDS

#extraccion y lectura con Rapids

```
start = time.time()
gdf = cudf.read_csv('df_data.csv', names='uno')
end = time.time()
```

```
print('Descripción de salida del dataset:')
gdf
```

Descripción de salida del dataset:

		u	n	o
0	"/edx/var/log/tracking/tracking.log-20171110-...	null	null	
1	/edx/var/log/tracking/tracking.log-20171110-15...	null	null	
2	/edx/var/log/tracking/tracking.log-20171110-15...	null	null	
3	/edx/var/log/tracking/tracking.log-20171110-15...	null	null	
4	/edx/var/log/tracking/tracking.log-20171110-15...	null	null	
...
79995	/edx/var/log/tracking/tracking.log-20171220-15...	null	null	
79996	/edx/var/log/tracking/tracking.log-20171220-15...	null	null	
79997	/edx/var/log/tracking/tracking.log-20171220-15...	null	null	
79998	/edx/var/log/tracking/tracking.log-20171220-15...	null	null	
79999	/edx/var/log/tracking/tracking.log-20171220-15...	null	null	

80000 rows × 3 columns

Extracción y Lectura

cuDF

`cudf.read_csv`

Aspecto familiar para los científicos de datos que trabajan en Python.

viernes 17 de julio d
e 2020

Proceso ETL: Transformación

```
#Creo DataFrame con Pandas
pdPandas = pd.DataFrame(a)
```



```
#convieto para cudf dataframe
```

```
gdf_new = cudf.DataFrame.from_pandas(pdPandas)
```

```
gdf_new.dropna(axis=0, how='any')
```

```
gdf_new.drop_column('accept_language') #columna si relevancia
gdf_new.drop_column('name') #muchos datos faltantes
```

```
context = gdf_new['context']
context.str.normalize_spaces()
newcontext = context.str.split(pat = ",")
```

```
newcontext.columns = ['a', 'b', 'c']
```

```
newData = gdf_new.join(newcontext, how='left')
newData.drop_column('context')
```

```
newData.sort_values('username')
newData.drop_duplicates(subset='username', keep=False, inplace=False, ignore_index=False)
```

```
newData = newData.reset_index()
```

RAPIDS

cuDF

cudf.DataFrame.from_pandas	Convierte de un dataframe de pandas a un dataframe cudf
dropna	Eliminar filas con datos perdidos.
drop_column	Eliminar columna seleccionada.
str.normalize_spaces	Normaliza espacios en blanco.
str.split	Dividir una columna y la incorpora como un nuevo dataframe.
columns	Crea columnas.
join	Unir dataframes.
drop_duplicates	Eliminar duplicados.
reset_index	Resetear el index.

	username	event_source	time	host	referer	ip	a	b	c
0	SheilaCarrion	server	2017-12-12T04:00:46.043766+00:00	opencampus.utpl.edu.ec	http://opencampus.utpl.edu.ec/courses/course-v...	190.152.211.115	{'course_user_tags': {}	'user_id': 13194	'org_id': 'UTPL'
1	SheilaCarrion	browser	2017-12-12T04:00:46.575445+00:00	opencampus.utpl.edu.ec	http://opencampus.utpl.edu.ec/courses/course-v...	190.152.211.115	{'user_id': 13194	'org_id': 'UTPL'	'course_id': 'course-v1:UTPL+EFHE-Ed7+2017_DIC'
2	SheilaCarrion	server	2017-12-12T04:00:46.644949+00:00	opencampus.utpl.edu.ec	http://opencampus.utpl.edu.ec/courses/course-v...	190.152.211.115	{'user_id': 13194	'org_id': "	'course_id': "
3	SheilaCarrion	server	2017-12-12T04:00:48.565070+00:00	opencampus.utpl.edu.ec	http://opencampus.utpl.edu.ec/courses/course-v...	190.152.211.115	{'course_user_tags': {}	'user_id': 13194	'org_id': 'UTPL'
4	SheilaCarrion	browser	2017-12-12T04:00:49.053330+00:00	opencampus.utpl.edu.ec	http://opencampus.utpl.edu.ec/courses/course-v...	190.152.211.115	{'user_id': 13194	'org_id': 'UTPL'	'course_id': 'course-v1:UTPL+EFHE-Ed7+2017_DIC'

Proceso ETL: Carga.

RAPIDS

```
da_data = 'data_mejora.csv'  
newData.to_csv(da_data)
```



data_mejora.csv

Carga

cuDF

	username	event_source	time	host	referer	ip	a	b	c
0	SheilaCarrion	server	2T04:00:46.043766+00:00	opencampus.utpl.edu.ec:d4c03ac21f6eabe317d7b/		190.152.211.115	{'course_user_tags': []}	'user_id': 13194	'org_id': 'UTPL'
1	SheilaCarrion	browser	2T04:00:46.575445+00:00	opencampus.utpl.edu.ec:d4c03ac21f6eabe317d7b/		190.152.211.115	{'user_id': 13194}	'org_id': 'UTPL' PL+EFHE-Ed7+2017_DIC	
2	SheilaCarrion	server	2T04:00:46.64949+00:00	opencampus.utpl.edu.ec:3a4a53b8739fc88fa055e7/		190.152.211.115	{'user_id': 13194}	'org_id': ''	'course_id': ''
3	SheilaCarrion	server	2T04:00:48.555070+00:00	opencampus.utpl.edu.ec:3a4a53b8739fc88fa055e7/		190.152.211.115	{'course_user_tags': []}	'user_id': 13194	'org_id': 'UTPL'
4	SheilaCarrion	browser	2T04:00:49.053330+00:00	opencampus.utpl.edu.ec:3a4a53b8739fc88fa055e7/		190.152.211.115	{'user_id': 13194}	'org_id': 'UTPL' PL+EFHE-Ed7+2017_DIC	
5	SheilaCarrion	server	2T04:00:49.129556+00:00	opencampus.utpl.edu.ec:d4c03ac21f6eabe317d7b/		190.152.211.115	{'user_id': 13194}	'org_id': ''	'course_id': ''
6	Amada1012	browser	2T04:00:59.155244+00:00	opencampus.utpl.edu.ec:54648a6078348ea41557/		190.57.140.77	{'user_id': 11041}	'org_id': 'UTPL' PL+EFHE-Ed7+2017_DIC	
7	Amada1012	server	2T04:00:59.219022+00:00	opencampus.utpl.edu.ec:54648a6078348ea41557/		190.57.140.77	{'course_user_tags': []}	'user_id': 11041	'org_id': 'UTPL'
8	Amada1012	server	2T04:00:59.323829+00:00	opencampus.utpl.edu.ec:54648a6078348ea41557/		190.57.140.77	{'course_user_tags': []}	'user_id': 11041	'org_id': 'UTPL'
9	Amada1012	browser	2T04:00:59.946183+00:00	opencampus.utpl.edu.ec:54648a6078348ea41557/		190.57.140.77	{'user_id': 11041}	'org_id': 'UTPL' PL+EFHE-Ed7+2017_DIC	
10	andilanas	server	2T04:01:01.432044+00:00	opencampus.utpl.edu.ec		186.4.129.70	{'course_user_tags': []}	'user_id': 10929	'org_id': 'UTPL'
11	andilanas	server	2T04:01:01.687954+00:00	opencampus.utpl.edu.ec:EFHE-Ed7+2017_DICinfo		186.4.129.70	{'user_id': 10929}	'org_id': ''	'course_id': ''
12	andilanas	server	2T04:01:08.200766+00:00	opencampus.utpl.edu.ec:pus.utpl.edu.ec/dashboard		186.4.129.70	{'course_user_tags': []}	'user_id': 10929	'org_id': 'UTPL'
13	andilanas	server	2T04:01:08.655762+00:00	opencampus.utpl.edu.ec:EFHE-Ed7+2017_DICinfo		186.4.129.70	{'user_id': 10929}	'org_id': ''	'course_id': ''
14	SheilaCarrion	server	2T04:03:16.818041+00:00	opencampus.utpl.edu.ec:d4c03ac21f6eabe317d7b/		190.152.211.115	{'course_user_tags': []}	'user_id': 13194	'org_id': 'UTPL'
15	SheilaCarrion	browser	2T04:03:17.342741+00:00	opencampus.utpl.edu.ec:d4c03ac21f6eabe317d7b/		190.152.211.115	{'user_id': 13194}	'org_id': 'UTPL' PL+EFHE-Ed7+2017_DIC	
16	SheilaCarrion	server	2T04:03:17.415023+00:00	opencampus.utpl.edu.ec:3a4a53b8739fc88fa055e7/		190.152.211.115	{'user_id': 13194}	'org_id': ''	'course_id': ''
17	SheilaCarrion	server	2T04:03:23.491310+00:00	opencampus.utpl.edu.ec:3a4a53b8739fc88fa055e7/		190.152.211.115	{'course_user_tags': []}	'user_id': 13194	'org_id': 'UTPL'
18	SheilaCarrion	browser	2T04:03:24.008458+00:00	opencampus.utpl.edu.ec:3a4a53b8739fc88fa055e7/		190.152.211.115	{'user_id': 13194}	'org_id': 'UTPL' PL+EFHE-Ed7+2017_DIC	
19	SheilaCarrion	server	2T04:03:24.082438+00:00	opencampus.utpl.edu.ec:34706b406a3ded0d22c15/		190.152.211.115	{'user_id': 13194}	'org_id': ''	'course_id': ''
20	SheilaCarrion	server	2T04:03:28.394948+00:00	opencampus.utpl.edu.ec:34706b406a3ded0d22c15/		190.152.211.115	{'course_user_tags': []}	'user_id': 13194	'org_id': 'UTPL'
21	SheilaCarrion	browser	2T04:03:28.921849+00:00	opencampus.utpl.edu.ec:34706b406a3ded0d22c15/		190.152.211.115	{'user_id': 13194}	'org_id': 'UTPL' PL+EFHE-Ed7+2017_DIC	
22	SheilaCarrion	server	2T04:03:29.000247+00:00	opencampus.utpl.edu.ec:184852ba304a88b0cfebad/		190.152.211.115	{'user_id': 13194}	'org_id': ''	'course_id': ''
23	SheilaCarrion	server	2T04:03:33.274316+00:00	opencampus.utpl.edu.ec:184852ba304a88b0cfebad/		190.152.211.115	{'course_user_tags': []}	'user_id': 13194	'org_id': 'UTPL'
24	SheilaCarrion	browser	2T04:03:33.997638+00:00	opencampus.utpl.edu.ec:184852ba304a88b0cfebad/		190.152.211.115	{'user_id': 13194}	'org_id': 'UTPL' PL+EFHE-Ed7+2017_DIC	
25	SheilaCarrion	server	2T04:03:34.077566+00:00	opencampus.utpl.edu.ec:64b1c953262bb08e91f09b/		190.152.211.115	{'user_id': 13194}	'org_id': ''	'course_id': ''

Proceso ML: Dividir dataset Train y Test

RAPIDS

```
from cuml.preprocessing.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(gdf_dataX, gdf_dataY, train_size=0.8)
```

**80% Train y
20% Test**

```
print(f'Original datasets: {gdf_dataX.shape[0], gdf_dataY.shape[0]} elements')  
print(f'Data X_train: {X_train.shape[0]} elements')  
print(f'Data X_test: {X_test.shape[0]} elements')  
print(f'Data y_train: {y_train.shape[0]} elements')  
print(f'Data y_test: {y_test.shape[0]} elements')
```

Dataset Original: (77912, 77912) elementos
Data X_train: 62329 elementos
Data X_test: 15583 elementos
Data y_train: 62329 elementos
Data y_test: 15583 elementos

cuML	
Clase	Descripción
preprocessing.model_selection train_test_split	Dividir los datos en cuatro objetos intercalados.

Proceso ML: Preprocesamiento.

RAPIDS

```
from cuml.preprocessing.LabelEncoder import LabelEncoder
```

```
le = LabelEncoder()
```

```
value_usernameX = le.fit_transform(X_train.username)
value_ipX = le.fit_transform(X_train.ip)
```

Datos sin categorizar

```
X_train.username.head(3)
```

```
0      MILTON
1      gcjuma
2      Mayra_2017
Name: username, dtype: object
```

```
X_train.ip.head(3)
```

```
0    181.196.160.40
1    186.46.207.242
2    186.70.132.111
Name: ip, dtype: object
```

Datos categorizados

```
value usernameX.head(3)
```

```
0    215
1    371
2    237
dtype: int16
```

```
value ipX.head(3)
```

```
0      209
1      390
2      459
dtype: int16
```

cuML	
Clase	Descripción
Preprocessing.LabelEncoder	Permite categorizar datos.

Proceso ML: Algoritmos Machine Learning.

RAPIDS

```
from cuml.linear_model import LinearRegression
```

```
dataTest = cudf.DataFrame()
dataTest['usernameValueXtrain'] = cp.asarray(value_usernameX, dtype=cp.float32)
dataTest['ipValueXtrain'] = cp.asarray(value_ipX, dtype=cp.float32)
```

```
new_value = cp.asarray(value_usernameY, dtype=cp.float32)
dataY = cudf.Series(new_value)
```

```
lr = LinearRegression(fit_intercept = True, normalize = False,
                      algorithm = "eig")
reg = lr.fit(dataTest, dataY)
```

```
print("Coeficientes:")
print(reg.coef_)
print("Intercepción:", reg.intercept_)
print('Predicción del modelo:')
print(lr.predict(dataTest))
```

```
Coeficientes:
0    1.000000e+00
1    7.146533e-09
dtype: float32
Intercepción: 0.0
Predicción del modelo:
0      467.0
1      214.0
2      214.0
3      214.0
4      300.0
...
62324   214.0
62325   214.0
62326   214.0
62327   214.0
62328   100.0
Length: 62329, dtype: float32
```

```
from cuml.ensemble import RandomForestClassifier as cuRFC
```

```
dataTrainUSER = cp.asarray(value_usernameX, dtype=cp.float32)
dataTestUSER = cp.asarray(value_usernameY, dtype=cp.int32)
```

```
cuml_modelRDF = cuRFC(max_features=1.0,
                      n_bins=2,
                      n_estimators=2)
cuml_modelRDF.fit(dataTrainUSER, dataTestUSER)
```

```
print("Predicción del modelo: ")
print(cuml_predict)
```

```
Predicción del modelo:
[422 214 214 ... 214 214 214]
```

cuML

Regresión y Clasificación

linear_model LinearRegression	Algoritmo regresión lineal
Ensemble RandomForestClassifier	Algoritmo Random Forest

Comparativa de resultados de Aceleración: Rapids y Pandas

 pandas

Proceso ETL	Tiempo en Segundos	Desarrollo	# filas
Extracción	0.94	<pre>#exportación y lectura con pandas inicio = time.time() pd_Pandas = pd.read_csv('../df_data.csv', names='uno') fin = time.time() calculatedPD = fin - inicio print("Cálculo de extracción y lectura con Pandas = {}".format(calculatedPD)) Cálculo de extracción y lectura con Pandas = 0.9458587169647217</pre>	80.000
Transformación	3.04	<pre>print('Cálculo de transformación cpu-Pandas:', cpuPandas) Cálculo de transformación cpu-Pandas: 3.0437815189361572</pre>	80.000
Carga	0.90	<pre>start = time.time() da_data = 'data_mejora.csv' generalData.to_csv(da_data) end = time.time() calculated = end - start print("Cálculo de carga cpu-Pandas = {}".format(calculated)) Cálculo de carga cpu-Pandas = 0.9098856449127197</pre>	80.000

 RAPIDS

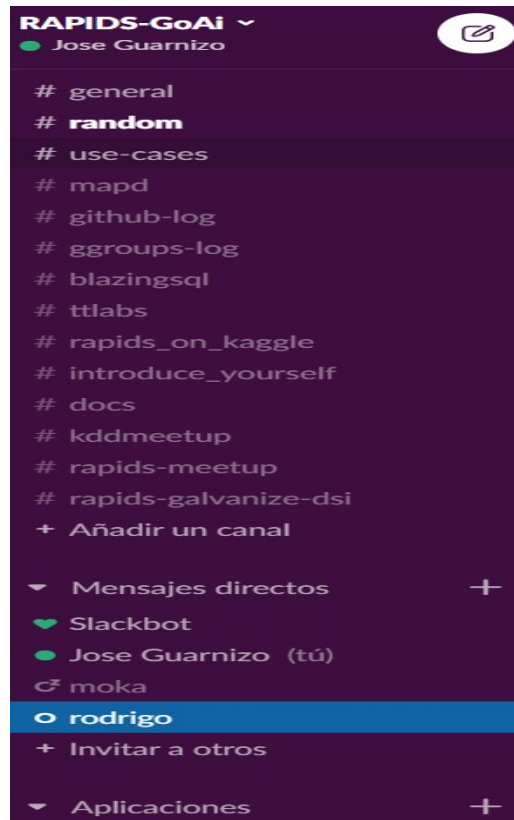
Proceso ETL	Tiempo en Segundos	Desarrollo	# filas
Extracción	0.41	<pre>start = time.time() gdf = cudf.read_csv('df_data.csv', names='uno') end = time.time() calculatedCUDF = end - start print("Cálculo de extracción y lectura con RAPIDS = {}".format(calculatedCUDF)) Cálculo de extracción y lectura con RAPIDS = 0.41393375396728516</pre>	80.000
Transformación	2.69	<pre>print('Cálculo de transformación gpu-RAPIDS:', calculatedGeneralRAPIDS) Cálculo de transformación gpu-RAPIDS: 2.696117401123047</pre>	80.000
Carga	0.37	<pre>import time start = time.time() da_data = 'data_mejora.csv' newData.to_csv(da_data) end = time.time() calculated = end - start print("Cálculo de carga gpu-Rapids = {}".format(calculated)) Cálculo de carga gpu-Rapids = 0.3737514019012451</pre>	80.000

ESTADO DEL TRABAJO DE TITULACIÓN

Avance del TT

Nombre Fase	% Avance Estimado	% Avance Real	% Retraso
Arquitectura RAPIDS	100%	100%	0%
Proceso de Instalación plataformas MERM	100%	100%	0%
Proceso ETL	100%	85%	15%
Proceso ML	100%	69%	31%

Detalles del proceso de Desarrollo



Jose Guarnizo 09:25

Hi guys, when installing RAPIDS in google colab i have problems with version 0.14 when importing cudf, can you help me please.



12 respuestas Última respuesta hace 21 días



rodrigo hace 21 días

hey [@Jose Guarnizo](#) you should consider app.blazingsql.com You get the same Tesla T4 GPU for free, it's vanilla JupyterLab, but all the RAPIDS packages are pre-installed and ready to go. No setup whatsoever. You can run the stable cuDF 0.14 or the nightly cuDF 0.15 inside the app. (editado)



Jose Guarnizo hace 21 días

Thanks [@rodrigo](#), i will try with blazingsql, Thank you so much for everything.



Taurean hace 21 días

[@Jose Guarnizo](#) can you print your output from the install cell? 2 things could be happening

1. you got a K80
2. the installation failed somewhere.

(editado)



Taurean hace 21 días

[@Jose Guarnizo](#) i see it....someone changed their files from what's on the yml created earlier this week and it failed to solve. I've reverted back to how we solved before and will look into `conda-pack`. It works as expected now. Thanks for bringing this to our attention!

Rapids

RAPIDS

[HOME](#) [ABOUT](#) [GET STARTED](#) [COMMUNITY](#) [BLOG](#) [DOCS](#) [GITHUB](#)

RAPIDS

Open GPU Data Science

[GET STARTED](#)

GPU DATA SCIENCE

⚡ ACCELERATED DATA SCIENCE

The RAPIDS suite of open source software libraries gives you the freedom to execute end-to-end data science and analytics pipelines entirely on GPUs.

🔗 SCALE OUT ON GPUS

Seamlessly scale from GPU workstations to multi-GPU servers and multi-node clusters with Dask.

[Learn about Dask >>](#)

🔌 PYTHON INTEGRATION

Accelerate your Python data science toolchain with minimal code changes and no new tools to learn.

[Learn about our libraries >>](#)

RAPIDS NEWS



📄 NATURAL LANGUAGE PROCESSING: TEXT PREPROCESSING AND VECTORIZING AT ROCKING SPEED WITH RAPIDS CUML

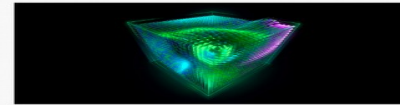
[post by Simon Andersen](#)



📄 TACKLING LARGE GRAPHS WITH RAPIDS CUGRAPH AND CUDA UNIFIED MEMORY ON GPUS

Out of Memory, @#\$%! The Out of Memory (OOM) error has to be one of the most frustrating errors to encounter. Unfortunately, when analyzing large graphs on GPUs using RAPIDS cuGraph, it's one of the errors most often encountered. Migrating...

[post by Alex Fender](#)



📄 RUN YOUR PYTHON USER DEFINED FUNCTIONS IN NATIVE CUDA KERNELS WITH RAPIDS CUDF

[post by Jiqun Tu](#)

🐦 JOHN MURRAY

RT @MurrayData: A time consuming task, in #datascience, is exploratory analysis of new data sources to understand field contents and relati...

[retweet by @rapidsai](#)

🐦 RAPIDS AI

@devnulling @adamlikesai @numba_jit @CuPy_Team My bad - here's the correct link <https://t.co/XdlW8PaVnr>

[post by @rapidsai](#)

🐦 RAPIDS AI

Exciting to see @rapidsai cuML is expanding support for #NLP! GPU-accelerated NLP transformers now perform blazing... <https://t.co/jK1jOG00yB>

[post by @rapidsai](#)

GRACIAS

¿PREGUNTAS?