

## MANUAL CURSO DE AGVD

### Página Oficial del Curso:

<https://jdvelasq.github.io/courses/analitica-de-grandes-datos/>

Todo el material ha sido construido por Juan David Velásquez, PhD, profesor asociado de la Universidad Nacional de Colombia para uso exclusivo de las clases de Analítica al interior del campus.

### Laboratorios:

<https://jdvelasq.github.io/courses/analitica-de-grandes-datos/grades.html>

\* Para poder acceder debe estar ingresado en el listado, si por algún motivo no esta por favor reportarlo en Slack. Actualmente esta en proceso de carga por lo que favor revisar su usuario el día 30 de noviembre de 2019.

### Página Slack:

La página oficial del curso es: [analiticagvdunalmed.slack.com](https://analiticagvdunalmed.slack.com)

### Github con códigos:

<https://github.com/amartinUnal/AnaliticaGVD/>

En este se colocarán los códigos que se utilizarán en clase, estos corresponden a los mismos códigos de la Página Oficial del Curso con algunos ejemplos adicionales.

### Instalación y configuración de herramientas para la clase:

1. Crear una cuenta en github:  
<https://jdvelasq.github.io/courses/analitica-de-grandes-datos/setup.html>
2. Descargar e instalar VirtualBox + Vagrant específico para Linux (caso de que utilices Windows):  
<https://github.com/jdvelasq/vagrant4analytics>

En caso de utilizar iOS / Linux se puede utilizar Docker directamente, para lo cual descargar e instalar Docker.

3. Para acceder a Vagrant seguir las instrucciones del repositorio:  
<https://jdvelasq.github.io/courses/analitica-de-grandes-datos/setup.html>

\* En caso de tener problemas con el disksize al momento de usar Vagrant, instalar el paquete con la siguiente instrucción en línea de comandos:

```
vagrant plugin install vagrant-disksize
```

```

Administrator: Command Prompt
Microsoft Windows [Version 10.0.17763.864]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd C:\Users\██████████\docker4analytics

C:\Users\██████████\docker4analytics>vagrant up
Bringing machine 'default' up with 'virtualbox' provider...
==> default: Checking if box 'ubuntu/bionic64' version '20190918.0.0' is up to date...
==> default: A newer version of the box 'ubuntu/bionic64' for provider 'virtualbox' is
==> default: available! You currently have version '20190918.0.0'. The latest is version
==> default: '20191125.0.0'. Run 'vagrant box update' to update.
==> default: Clearing any previously set forwarded ports...
==> default: Clearing any previously set network interfaces...
==> default: Preparing network interfaces based on configuration...
    default: Adapter 1: nat
==> default: Forwarding ports...
    default: 3333 (guest) => 3333 (host) (adapter 1)
    default: 8888 (guest) => 8888 (host) (adapter 1)
    default: 5000 (guest) => 5000 (host) (adapter 1)
    default: 8000 (guest) => 8000 (host) (adapter 1)
    default: 6006 (guest) => 6006 (host) (adapter 1)
    default: 9000 (guest) => 9000 (host) (adapter 1)
    default: 50070 (guest) => 50070 (host) (adapter 1)
    default: 8089 (guest) => 8088 (host) (adapter 1)
    default: 8080 (guest) => 8080 (host) (adapter 1)
    default: 8881 (guest) => 8881 (host) (adapter 1)
    default: 8880 (guest) => 8880 (host) (adapter 1)
    default: 4040 (guest) => 4040 (host) (adapter 1)
    default: 4041 (guest) => 4041 (host) (adapter 1)
    default: 3088 (guest) => 3088 (host) (adapter 1)
    default: 22 (guest) => 2222 (host) (adapter 1)
==> default: Running 'pre-boot' VM customizations...
==> default: Booting VM...
==> default: Waiting for machine to boot. This may take a few minutes...
    default: SSH address: 127.0.0.1:2222
    default: SSH username: vagrant
    default: SSH auth method: private key
==> default: Machine booted and ready!
==> default: Checking for guest additions in VM...
    default: The guest additions on this VM do not match the installed version of
    default: VirtualBox! In most cases this is fine, but in rare cases it can
    default: prevent things such as shared folders from working properly. If you see
    default: shared folder errors, please make sure the guest additions within the
    default: virtual machine match the version of VirtualBox you have installed on
    default: your host and reload your VM.
    default:
    default: Guest Additions Version: 5.2.32
    default: VirtualBox Version: 6.0
==> default: Mounting shared folders...
    default: /vagrant => C:/Users/amant/UN/AnaliticaGVD/docker4analytics
==> default: Machine already provisioned. Run 'vagrant provision' or use the '--provision'
==> default: flag to force provisioning. Provisioners marked to run always will still run.

C:\Users\██████████\docker4analytics>vagrant ssh_

```

```

vagrant@ubuntu-bionic: /vagrant

C:\Users\██████████\docker4analytics>vagrant ssh
Welcome to Ubuntu 18.04.3 LTS (GNU/Linux 4.15.0-70-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

System information as of Wed Nov 27 19:56:35 UTC 2019

System load:  0.27           Processes:    120
Usage of /:   36.4% of 14.48GB Users logged in:   0
Welcome to Ubuntu 18.04.3 LTS (GNU/Linux 4.15.0-70-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

System information as of Wed Nov 27 19:56:35 UTC 2019

System load:  0.27           Processes:    120
Usage of /:   36.4% of 14.48GB Users logged in:   0
Memory usage: 1%            IP address for enp0s3: 10.0.2.15
Swap usage:   0%            IP address for docker0: 172.17.0.1

 * Overheard at KubeCon: "microk8s.status just blew my mind".

https://microk8s.io/docs/commands#microk8s.status

27 packages can be updated.
0 updates are security updates.

Last login: Wed Nov 27 13:30:06 2019 from 10.0.2.2
vagrant@ubuntu-bionic:~$ cd /vagrant
vagrant@ubuntu-bionic:/vagrant$

```

4. Si tienes Linux / iOS corres el Docker descargando la versión específica correspondiente al bloque de trabajo del curso.

- Bloque 1: MapReduce
- Bloque 2: Pig
- Bloque 3: Hive
- Bloque 4: Spark

Con la instrucción:

```
docker run --rm -it -v "$PWD":/datalake --name hive -p 50070:50070 -p 8088:8088 -p 8888:8888 -p 5000:5000 jdvelasq/<nombre del contenedor>
```

\* Si estas usando el `vagrant` con SO Ubuntu y te genera error de permisos, recuerda adicionar el `sudo` antes del `docker`.

\* La versión pseudo monta `hdfs`, mientras que la versión `standalone` trabaja con el sistema de archivos local de la máquina. Por lo anterior la versión `standalone` puede ser de ejecución más rápida, pero la pseudo aprovecha el sistema de almacenamiento `hdfs` estándar cuando se trabaja en `hadoop`.

\* Recuerda correrlo la primera vez en una red distinta a la de la universidad dado que se requiere acceso para descargar ciertos componentes como Docker, las cuales están restringidas al interior del campus. Luego de la primera ejecución los archivos quedan guardados en tu maquina y ya es posible ejecutarlo en la red del campus.

## 5. Bloque 1: MapReduce

Para la primera clase vamos a utilizar: <https://colab.research.google.com>

```
docker run --rm -it -v "$PWD":/datalake --name hadoop -p 50070:50070 -p 8088:8088 -p 8888:8888 -p 5000:5000 jdvelasq/hadoop:2.8.5-pseudo
```

```
root@9b3936be3c8c:/datalake
vagrant@ubuntu-bionic:/vagrant$ docker run --rm -it -v "$PWD":/datalake --name hadoop -p 50070:50070 -p 8088:8088 -p 8888:8888 -p 5000:5000 jdvelasq/hadoop:2.8.5-pseudo
* Starting OpenBSD Secure Shell server sshd
[ OK ]
Could not load host key: /etc/ssh/ssh_host_rsa_key
Formatting using clusterid: CID-c7b0dc1a-4efb-4010-84d6-124e4b098a2d
Starting namenodes on [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-9b3936be3c8c.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-9b3936be3c8c.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-9b3936be3c8c.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-9b3936be3c8c.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-9b3936be3c8c.out

-----
Hadoop NameNode at:
    http://127.0.0.1:50070/
Yarn ResourceManager at:
    http://127.0.0.1:8088/
-----
root@9b3936be3c8c:/datalake#
```

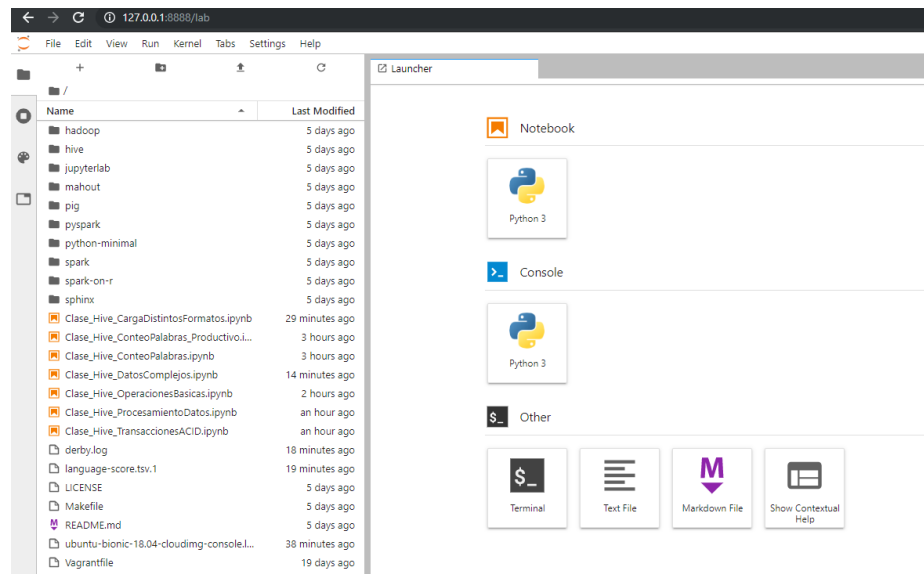
\* Para validar el funcionamiento de los nodos y del Yarn se puede ingresar al navegador la url 127.0.0.1 con el respectivo puerto

En este caso ya esta corriendo el docker para Hadoop, basta digitar Python o si se desea trabajar en Jupyter basta colocar la instrucción: `jupyter lab --ip=0.0.0.0`

```
Select root@9b3936be3c8c: /datalake
root@9b3936be3c8c:/datalake# jupyter lab --ip=0.0.0.0
[I 23:07:31.456 LabApp] Writing notebook server cookie secret to /root/.local/share/jupyter/runtime/notebook_cookie_secret
[I 23:07:31.758 LabApp] JupyterLab extension loaded from /usr/local/lib/python3.6/dist-packages/jupyterlab
[I 23:07:31.758 LabApp] JupyterLab application directory is /usr/local/share/jupyter/lab
[I 23:07:31.922 LabApp] Serving notebooks from local directory: /datalake
[I 23:07:31.922 LabApp] The Jupyter Notebook is running at:
[I 23:07:31.922 LabApp] http://9b3936be3c8c:8888/?token=38e936cf45d66df4c6cd26f2f8e4eb8cddc3959e2e786632
[I 23:07:31.922 LabApp] or http://127.0.0.1:8888/?token=38e936cf45d66df4c6cd26f2f8e4eb8cddc3959e2e786632
[I 23:07:31.922 LabApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 23:07:31.926 LabApp]

To access the notebook, open this file in a browser:
file:///root/.local/share/jupyter/runtime/nbserver-1280-open.html
Or copy and paste one of these URLs:
http://9b3936be3c8c:8888/?token=38e936cf45d66df4c6cd26f2f8e4eb8cddc3959e2e786632
or http://127.0.0.1:8888/?token=38e936cf45d66df4c6cd26f2f8e4eb8cddc3959e2e786632
```

Selecciona la url y la pegas en el explorador:



Los ejemplos de clase corresponderán a:

- <https://jdvelasq.github.io/courses/notebooks/hadoop/1-01-intro-mapreduce.html>
- <https://jdvelasq.github.io/courses/notebooks/hadoop/1-03-wordcount-streaming-python.html>
- <https://jdvelasq.github.io/courses/notebooks/hadoop/1-04-wordcount-python-efficient.html>
- <https://jdvelasq.github.io/courses/notebooks/hadoop/1-07-wordcount-standalone.html>
- [https://jdvelasq.github.io/courses/notebooks/hadoop/1-08-\(opcional\)-wordcount-java-standalone.html](https://jdvelasq.github.io/courses/notebooks/hadoop/1-08-(opcional)-wordcount-java-standalone.html)

- [https://jdvelasq.github.io/courses/notebooks/hadoop/1-09-\(opcional\)-wordcount-java-pseudo.html](https://jdvelasq.github.io/courses/notebooks/hadoop/1-09-(opcional)-wordcount-java-pseudo.html)

ejemplo:

```
Clase_HadoopMR_Concep
+ %K 
Code
[1]: ## Archivos de prueba
## A continuación se generarán tres archivos de prueba para probar el sistema. Puede usar directamente comandos del sistema operativo en el Terminal y
## el editor de texto pico para crear los archivos.

[2]: ## Se crea el directorio de entrada
rm -rf input output
mkdir input

[3]: %writefile input/text0.txt
Analytics is the discovery, interpretation, and communication of meaningful patterns
in data. Especially valuable in areas rich with recorded information, analytics relies
on the simultaneous application of statistics, computer programming and operations research
to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business
performance. Specifically, areas within analytics include predictive analytics, prescriptive
analytics, enterprise decision management, descriptive analytics, cognitive analytics, Big
Data Analytics, retail analytics, store assortment and stock-keeping unit optimization,
marketing optimization and marketing mix modeling, web analytics, call analytics, speech
analytics, sales force sizing and optimization, price and promotion modeling, predictive
science, credit risk analysis, and fraud analytics. Since analytics can require extensive
computation (see big data), the algorithms and software used for analytics harness the most
current methods in computer science, statistics, and mathematics.

Writing input/text0.txt

[4]: %writefile input/text1.txt
The field of data analysis. Analytics often involves studying past historical data to
research potential trends, to analyze the effects of certain decisions or events, or to
evaluate the performance of a given tool or scenario. The goal of analytics is to improve
the business by gaining knowledge which can be used to make improvements or changes.

Writing input/text1.txt

[5]: %writefile input/text2.txt
Data analytics (DA) is the process of examining data sets in order to draw conclusions
about the information they contain, increasingly with the aid of specialized systems
and software. Data analytics technologies and techniques are widely used in commercial
industries to enable organizations to make more-informed business decisions and by
scientists and researchers to verify or disprove scientific models, theories and
hypotheses.

Writing input/text2.txt
```

En caso de preferir por consola puedes utilizar el editor `pico`.

Los códigos están en [github/AnaliticaGVD](https://github.com/amartinUnal/AnaliticaGVD) con sus respectivos resultados de ejecución en la estructura de carpetas por nombre:

**amartinUnal / AnaliticaGVD**

Unwatch 1
Star 0
Fork 0

Code
Issues 0
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
Settings

Branch: master
**AnaliticaGVD / CodigoClase / 1-MapReduce /**

Create new file
Upload files
Find file
History

**amartinUnal** Add files via upload
Latest commit 608064b 20 seconds ago

..		
01-intro-mapreduce	Add files via upload	3 minutes ago
03-wordcount-streaming-python	Add files via upload	3 minutes ago
04-wordcount-python-efficient	Add files via upload	3 minutes ago
OP-Transacciones	Add files via upload	20 seconds ago

## Archivo de conteo de palabras:

```
In [1]: ## Archivos de prueba
## A continuación se generarán tres archivos de prueba para probar el sistema. Puede usar directamente c
omandos del sistema operativo en el Terminal y
## el editor de texto pico para crear los archivos.

In [2]: ## Se crea el directorio de entrada
rm -rf input output
mkdir input

In [3]: %writefile input/text0.txt
Analytics is the discovery, interpretation, and communication of meaningful patterns
in data. Especially valuable in areas rich with recorded information, analytics relies
on the simultaneous application of statistics, computer programming and operations research
to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business
performance. Specifically, areas within analytics include predictive analytics, prescriptive
analytics, enterprise decision management, descriptive analytics, cognitive analytics, Big
Data Analytics, retail analytics, store assortment and stock-keeping unit optimization,
marketing optimization and marketing mix modeling, web analytics, call analytics, speech
analytics, sales force sizing and optimization, price and promotion modeling, predictive
science, credit risk analysis, and fraud analytics. Since analytics can require extensive
computation (see big data), the algorithms and software used for analytics harness the most
current methods in computer science, statistics, and mathematics.

Writing input/text0.txt
```

## 6. Bloque 2: Pig

```
docker run --rm -it -v "$PWD":/datalake --name pig -p 50070:50070 -p 8088:8088
-p 8888:8888 -p 5000:5000 jdvelasq/pig:0.17.0-pseudo
```

```
root@3c409129cda1:/datalake
vagrant@ubuntu-bionic:/vagrant$ docker run --rm -it -v "$PWD":/datalake --name pig -p 50070:50070 -p 8088:8088 -p 8888:8888 -p 5000:5000 jdvelasq/pig:0.17.0-pseudo
* Starting OpenBSD Secure Shell server sshd
[ OK ]
Formatting using clusterid: CID-5187c0f3-56d5-4531-a7ec-06bfd76e03a2
Starting namenodes on [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-3c409129cda1.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-3c409129cda1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-3c409129cda1.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-3c409129cda1.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-3c409129cda1.out

Hadoop NameNode at:
http://127.0.0.1:50070/

Yarn ResourceManager at:
http://127.0.0.1:8088/

starting historyserver, logging to /usr/local/hadoop/logs/mapred--historyserver-3c409129cda1.out
root@3c409129cda1:/datalake#
```

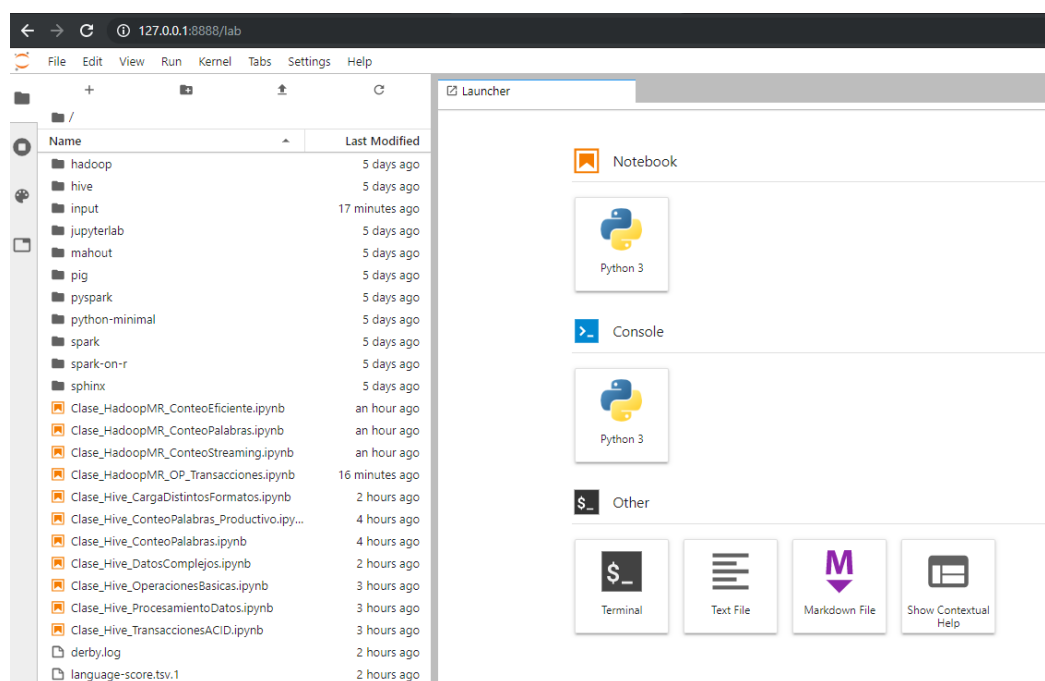
\* Para validar el funcionamiento de los nodos y del Yarn se puede ingresar al navegador la url 127.0.0.1 con el respectivo puerto

En este caso ya esta corriendo el docker para pig, basta con escribir pig o si se desea trabajar en Jupyter basta colocar la instrucción: `jupyter lab --ip=0.0.0.0`

```
Select root@3c409129cda1:/datalake
root@3c409129cda1:/datalake# jupyter lab --ip=0.0.0.0
[I 00:35:18.298 LabApp] Writing notebook server cookie secret to /root/.local/share/jupyter/runtime/notebook_cookie_secret
[I 00:35:19.120 LabApp] JupyterLab extension loaded from /usr/local/lib/python3.6/dist-packages/jupyterlab
[I 00:35:19.120 LabApp] JupyterLab application directory is /usr/local/share/jupyter/lab
[I 00:35:19.557 LabApp] Serving notebooks from local directory: /datalake
[I 00:35:19.558 LabApp] The Jupyter Notebook is running at:
[I 00:35:19.558 LabApp] http://3c409129cda1:8888/?token=f490ec29e91d86bd7146c3db510d893c75d2e9d71209320d
[I 00:35:19.558 LabApp] or http://127.0.0.1:8888/?token=f490ec29e91d86bd7146c3db510d893c75d2e9d71209320d
[I 00:35:19.558 LabApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 00:35:19.569 LabApp]

To access the notebook, open this file in a browser:
file:///root/.local/share/jupyter/runtime/nbserver-1368-open.html
Or copy and paste one of these URLs:
http://3c409129cda1:8888/?token=f490ec29e91d86bd7146c3db510d893c75d2e9d71209320d
or http://127.0.0.1:8888/?token=f490ec29e91d86bd7146c3db510d893c75d2e9d71209320d
```

Selecciona la url y la pegas en el explorador:



Los ejemplos de clase corresponderán a:

- <https://jdvelasq.github.io/courses/notebooks/pig/1-01-wordcount-pig.html>
- <https://jdvelasq.github.io/courses/notebooks/pig/1-02-pig-basics.html>
- <https://jdvelasq.github.io/courses/notebooks/pig/2-04-advanced.html>
- <https://jdvelasq.github.io/courses/notebooks/pig/2-05-tipos-de-datos.html>
- <https://jdvelasq.github.io/courses/notebooks/pig/2-06-udfs.html>
- <https://jdvelasq.github.io/courses/notebooks/pig/2-07-standalone.html>

ejemplo:

```
Clase_Pig_Wordcount.ipynb
+ ✕ 📄 📁 📄 Code
[1]: ## Archivos de prueba
## A continuación se generarán tres archivos de prueba para probar el sistema. Puede usar directamente comandos del sistema operativo en el Terminal y el editor de
## texto pico para crear los archivos.

[2]: ## Se crea el directorio de entrada
rm -rf input output
mkdir input

[3]: %writefile input/text0.txt
Analytics is the discovery, interpretation, and communication of meaningful patterns
in data. Especially valuable in areas rich with recorded information, analytics relies
on the simultaneous application of statistics, computer programming and operations research
to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business
performance. Specifically, areas within analytics include predictive analytics, prescriptive
analytics, enterprise decision management, descriptive analytics, cognitive analytics, Big
Data Analytics, retail analytics, store assortment and stock-keeping unit optimization,
marketing optimization and marketing mix modeling, web analytics, call analytics, speech
analytics, sales force sizing and optimization, price and promotion modeling, predictive
science, credit risk analysis, and fraud analytics. Since analytics can require extensive
computation (see big data), the algorithms and software used for analytics harness the most
current methods in computer science, statistics, and mathematics.

Writing input/text0.txt

[4]: %writefile input/text1.txt
The field of data analysis, analytics often involves studying past historical data to
research potential trends, to analyze the effects of certain decisions or events, or to
evaluate the performance of a given tool or scenario. The goal of analytics is to improve
the business by gaining knowledge which can be used to make improvements or changes.


Writing input/text1.txt

[5]: %writefile input/text2.txt
Data analytics (DA) is the process of examining data sets in order to draw conclusions
about the information they contain, increasingly with the aid of specialized systems
and software. Data analytics technologies and techniques are widely used in commercial
industries to enable organizations to make more informed business decisions and by
scientists and researchers to verify or disprove scientific models, theories and
hypotheses.

Writing input/text2.txt
```

En caso de preferir por consola puedes utilizar el editor `pico`.

Los códigos están en [github/AnaliticaGVD](https://github.com/AnaliticaGVD) con sus respectivos resultados de ejecución en la estructura de carpetas por nombre:

 **amartinUnal / AnaliticaGVD**

Code

Issues 0


Pull requests 0

Actions






Projects 0

Wiki

Branch: master ▾ **AnaliticaGVD / CodigosClase / 2-Pig /**

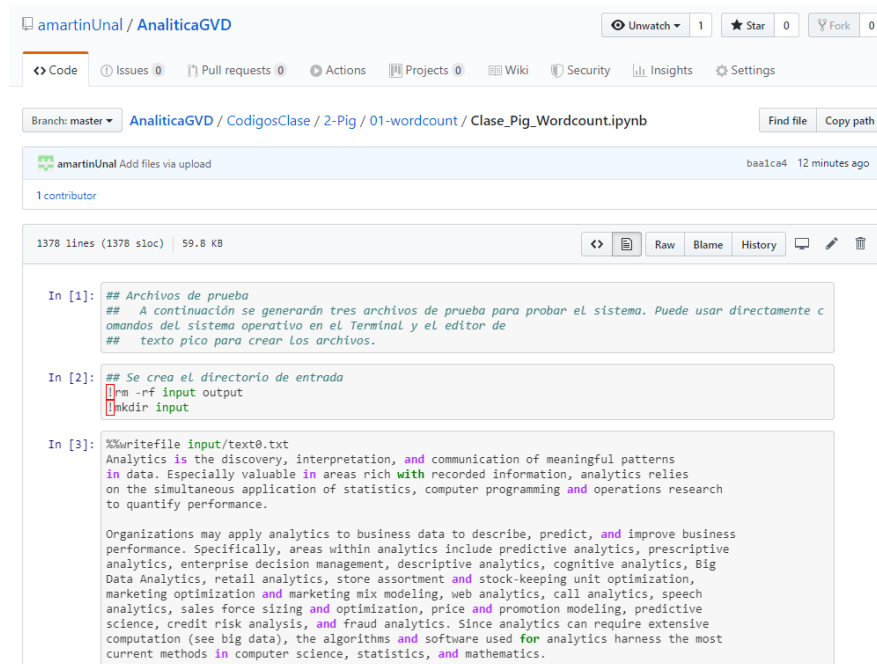
 **amartinUnal** Add files via upload

..

 <b>01-wordcount</b>	Add files via upload
 <b>02-pig-basics</b>	Add files via upload
 <b>04-advanced</b>	Add files via upload
 <b>05-tipos-de-datos</b>	Add files via upload
 <b>06-udfs</b>	Add files via upload



## Archivo de conteo de palabras:



```
## Archivos de prueba
## A continuación se generarán tres archivos de prueba para probar el sistema. Puede usar directamente c
omandos del sistema operativo en el Terminal y el editor de
## texto pico para crear los archivos.

## Se crea el directorio de entrada
rm -rf input output
mkdir input

%%writefile input/text0.txt
Analytics is the discovery, interpretation, and communication of meaningful patterns
in data. Especially valuable in areas rich with recorded information, analytics relies
on the simultaneous application of statistics, computer programming and operations research
to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business
performance. Specifically, areas within analytics include predictive analytics, prescriptive
analytics, enterprise decision management, descriptive analytics, cognitive analytics, Big
Data Analytics, retail analytics, store assortment and stock-keeping unit optimization,
marketing optimization and marketing mix modeling, web analytics, call analytics, speech
analytics, sales force sizing and optimization, price and promotion modeling, predictive
science, credit risk analysis, and fraud analytics. Since analytics can require extensive
computation (see big data), the algorithms and software used for analytics harness the most
current methods in computer science, statistics, and mathematics.
```

## 7. Bloque 3: Hive

```
docker run --rm -it -v "$PWD":/datalake --name hive -p 50070:50070 -p 8088:8088
-p 8888:8888 -p 5000:5000 jdvelasq/hive:2.3.6-pseudo
```

```
Select root@499bf3956267: /datalake
vagrant@ubuntu-bionic:~$ docker run --rm -it -v "$PWD":/datalake --name hive -p 50070:50070 -p 8088:8088 -p 8888:8888 -p 5000:5000 jdvelasq/hive:2.3.6-pseudo
* Starting OpenBSD Secure Shell server sshd
could not load host key: /etc/ssh/ssh_host_rsa_key

Formatting using clusterid: CID-f5191ab6-a394-4262-a120-c01b981b3177
Starting namenodes on [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-499bf3956267.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-499bf3956267.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-499bf3956267.out
starting yarn daemons
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-499bf3956267.out

Hadoop NameNode at:
    http://127.0.0.1:50070/

Yarn ResourceManager at:
    http://127.0.0.1:8088/

=====
mkdir: /user/: File exists
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Metastore connection URL:      jdbc:derby:;databaseName=/root/metastore_db;create=true
Metastore Connection Driver :  org.apache.derby.jdbc.EmbeddedDriver
Metastore connection User:    APP
Starting metastore schema initialization to 2.3.0
Initialization script hive-schema-2.3.0.derby.sql
Initialization script completed
schemaTool completed
starting historyserver, logging to /usr/local/hadoop/logs/mapred--historyserver-499bf3956267.out
```

\* Para validar el funcionamiento de los nodos y del Yarn se puede ingresar al navegador la url 127.0.0.1 con el respectivo puerto

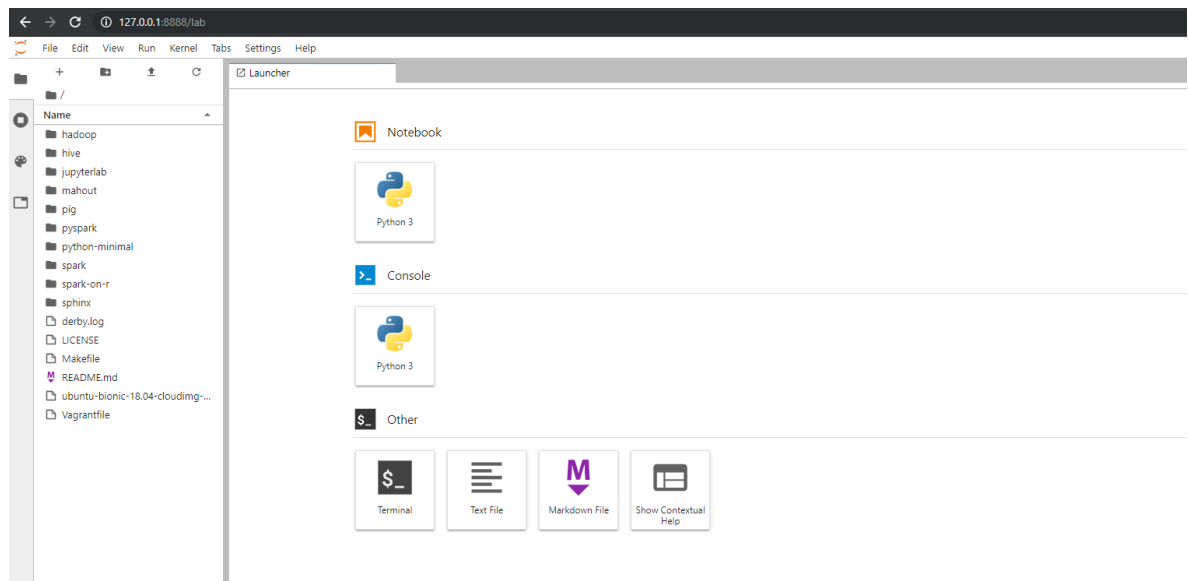
En este caso ya esta corriendo el docker para hive, basta con escribir hive o si se desea trabajar en Jupyter basta colocar la instrucción:

```
jupyter lab -ip=0.0.0.0
```

```
root@499bf3956267:/datalake# jupyter lab --ip=0.0.0.0
[I 20:05:38.191 LabApp] Writing notebook server cookie secret to /root/.local/share/jupyter/runtime/notebook_cookie_secret
[I 20:05:39.427 LabApp] JupyterLab extension loaded from /usr/local/lib/python3.6/dist-packages/jupyterlab
[I 20:05:39.428 LabApp] JupyterLab application directory is /usr/local/share/jupyter/lab
[I 20:05:39.883 LabApp] Serving notebooks from local directory: /datalake
[I 20:05:39.884 LabApp] The Jupyter Notebook is running at:
[I 20:05:39.884 LabApp] http://499bf3956267:8888/?token=8a362f08afa0c4eec7e79ee50494435de738a8c6c1926ec0
[I 20:05:39.884 LabApp] or http://127.0.0.1:8888/?token=8a362f08afa0c4eec7e79ee50494435de738a8c6c1926ec0
[I 20:05:39.884 LabApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 20:05:39.898 LabApp]

To access the notebook, open this file in a browser:
file:///root/.local/share/jupyter/runtime/nbserver-1733-open.html
Or copy and paste one of these URLs:
http://499bf3956267:8888/?token=8a362f08afa0c4eec7e79ee50494435de738a8c6c1926ec0
or http://127.0.0.1:8888/?token=8a362f08afa0c4eec7e79ee50494435de738a8c6c1926ec0
```

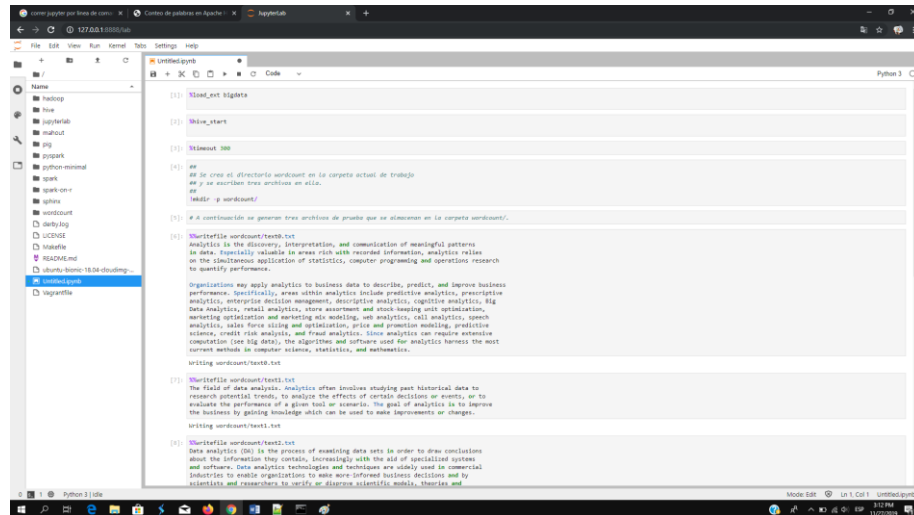
Selecciona la url y la pegas en el explorador:



Los ejemplos de clase corresponderán a:

- <https://jdvelasq.github.io/courses/notebooks/hive/1-01-conteo-de-palabras-en-hive.html>
- <https://jdvelasq.github.io/courses/notebooks/hive/1-02-operaciones-basicas-en-hive.html>
- <https://jdvelasq.github.io/courses/notebooks/hive/2-03-procesamiento-de-datos-con-hive.html>
- <https://jdvelasq.github.io/courses/notebooks/hive/2-07-manejo-de-datos-complejos.html>
- <https://jdvelasq.github.io/courses/notebooks/hive/2-08-standalone.html>

Ejemplo:



```
[1]: %load_ext bigdata

[2]: %hive_start

[3]: %timeout 300

[4]: ##
## Se crea el directorio wordcount en la carpeta actual de trabajo
## y se escriben tres archivos en ella.
##
%mkdir -p wordcount/

[5]: ## A continuación se generan tres archivos de prueba que se almacenan en la carpeta wordcount/.

[6]: %writefile wordcount/text0.txt
Analytics is the discovery, interpretation, and communication of meaningful patterns
in data. Especially valuable in areas rich with recorded information, analytics relies
on the simultaneous application of statistics, computer programming and operations research
to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business
performance. Specifically, areas within analytics include predictive analytics, prescriptive
analytics, enterprise decision management, descriptive analytics, cognitive analytics, Big
Data analytics, retail analytics, store assortment and stock-keeping unit optimization,
marketing optimization and marketing mix modeling, web analytics, call analytics, speech
analytics, sales force sizing and optimization, price and promotion modeling, predictive
science, credit risk analysis, and fraud analytics. Since analytics can require extensive
computation (see big data), the algorithms and software used for analytics harness the most
current methods in computer science, statistics, and mathematics.

Writing wordcount/text0.txt

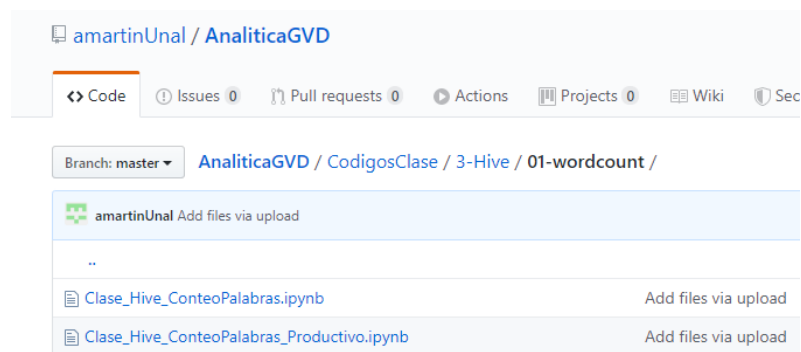
[7]: %writefile wordcount/text1.txt
The field of data analysis, analytics often involves studying past historical data to
research potential trends, to evaluate the effects of certain decisions or events, or to
evaluate the performance of a given tool or scenario. The goal of analytics is to improve
the business by gaining knowledge which can be used to make improvements or changes.

Writing wordcount/text1.txt

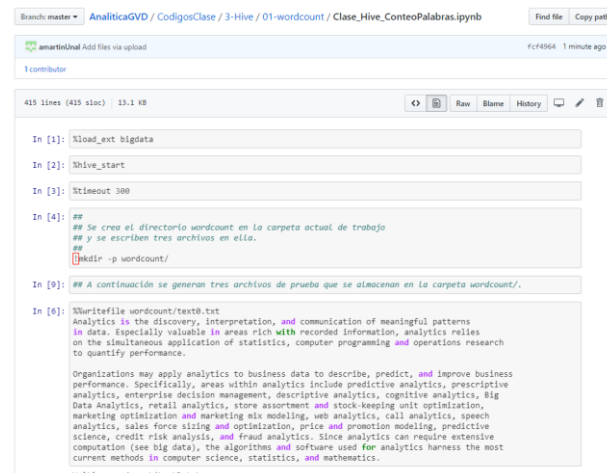
[8]: %writefile wordcount/text2.txt
Data analytics (DA) is the process of examining data sets in order to draw conclusions
about the information they contain, increasingly with the aid of specialized systems
and software. Data analytics technologies and techniques are widely used in commercial
industries to enable organizations to make more informed business decisions and by
scientists and researchers to verify or disprove scientific models, theories and
```

En caso de preferir por consola puedes utilizar el editor `pico`.

Los códigos están el [github/AnaliticaGVD](https://github.com/amartinUnal/AnaliticaGVD) con sus respectivos resultados de ejecución en la estructura de carpetas por nombre:



Archivo de conteo de palabras:



```
Branch: master AnaliticaGVD / CodigosClase / 3-Hive / 01-wordcount / Clase_Hive_ConteoPalabras.ipynb Find file Copy path

amartinUnal Add files via upload fcf4864 1 minute ago
1 contributor

415 lines (415 sloc) 13.1 KB Raw Blame History

In [1]: %load_ext bigdata

In [2]: %hive_start

In [3]: %timeout 300

In [4]: ##
## Se crea el directorio wordcount en la carpeta actual de trabajo
## y se escriben tres archivos en ella.
##
%mkdir -p wordcount/

In [9]: ## A continuación se generan tres archivos de prueba que se almacenan en la carpeta wordcount/.

In [6]: %writefile wordcount/text0.txt
Analytics is the discovery, interpretation, and communication of meaningful patterns
in data. Especially valuable in areas rich with recorded information, analytics relies
on the simultaneous application of statistics, computer programming and operations research
to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business
performance. Specifically, areas within analytics include predictive analytics, prescriptive
analytics, enterprise decision management, descriptive analytics, cognitive analytics, Big
Data analytics, retail analytics, store assortment and stock-keeping unit optimization,
marketing optimization and marketing mix modeling, web analytics, call analytics, speech
analytics, sales force sizing and optimization, price and promotion modeling, predictive
science, credit risk analysis, and fraud analytics. Since analytics can require extensive
computation (see big data), the algorithms and software used for analytics harness the most
current methods in computer science, statistics, and mathematics.

Writing wordcount/text0.txt
```

## 8. Bloque 4: Spark

```
docker run --rm -it -v "$PWD":/datalake --name pyspark -p 50070:50070 -p 8088:8088 -p 8888:8888 -p 5000:5000 jdvelasq/pyspark:2.4.4-pseudo
```

```
root@76028326a16e:/datalake
jvelasq@buntu:~/vagrant$ docker run --rm -it -v "$PWD":/datalake --name pyspark -p 50070:50070 -p 8088:8088 -p 8888:8888 -p 5000:5000 jdvelasq/pyspark:2.4.4-pseudo
* Starting OpenBSD Secure Shell server sshd
Could not load host key: /etc/ssh/ssh_host_rsa_key
[ OK ]
Formatting using clusterid: CID-eb8a593-6b74-4099-8663-5988b53e10ca
Starting namenodes on [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-76028326a16e.out
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-76028326a16e.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-76028326a16e.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-76028326a16e.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-76028326a16e.out

-----
Hadoop NameNode at:
    http://127.0.0.1:50070/
Yarn ResourceManager at:
    http://127.0.0.1:8088/
-----
root@76028326a16e:/datalake#
```

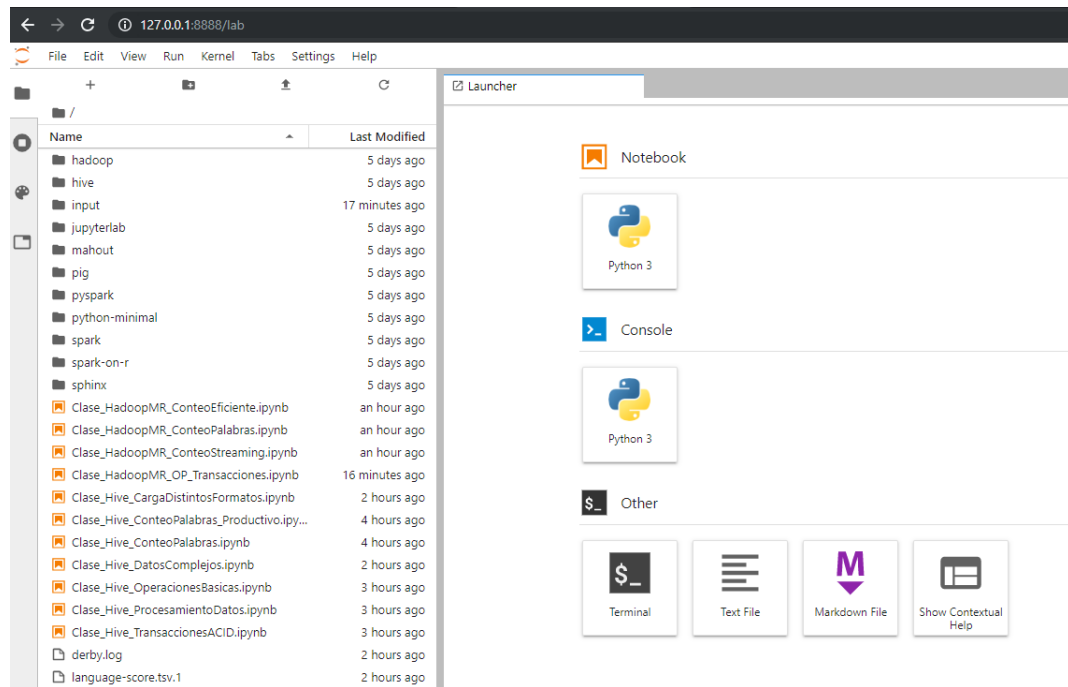
\* Para validar el funcionamiento de los nodos y del Yarn se puede ingresar al navegador la url 127.0.0.1 con el respectivo puerto

En este caso ya esta corriendo el docker para trabajar con spark, basta con escribir el código en un archivo y ejecutarlo por línea de comandos (o instrucción a instrucción en Python) o si se desea trabajar en Jupyter basta colocar la instrucción: `jupyter lab --ip=0.0.0.0`

```
root@3c409129cda1:/datalake# jupyter lab --ip=0.0.0.0
[I 00:35:18.298 LabApp] Writing notebook server cookie secret to /root/.local/share/jupyter/runtime/notebook_cookie_secret
[I 00:35:19.120 LabApp] JupyterLab extension loaded from /usr/local/lib/python3.6/dist-packages/jupyterlab
[I 00:35:19.120 LabApp] JupyterLab application directory is /usr/local/share/jupyter/lab
[I 00:35:19.557 LabApp] Serving notebooks from local directory: /datalake
[I 00:35:19.558 LabApp] The Jupyter Notebook is running at:
[I 00:35:19.558 LabApp] http://3c409129cda1:8888/?token=f490ec29e91d86bd7146c3db510d893c75d2e9d71209320d
[I 00:35:19.558 LabApp] or http://127.0.0.1:8888/?token=f490ec29e91d86bd7146c3db510d893c75d2e9d71209320d
[I 00:35:19.558 LabApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 00:35:19.569 LabApp]

To access the notebook, open this file in a browser:
file:///root/.local/share/jupyter/runtime/nbserver-1368-open.html
Or copy and paste one of these URLs:
http://3c409129cda1:8888/?token=f490ec29e91d86bd7146c3db510d893c75d2e9d71209320d
or http://127.0.0.1:8888/?token=f490ec29e91d86bd7146c3db510d893c75d2e9d71209320d
```

Selecciona la url y la pegas en el explorador:



Los ejemplos de clase corresponderán a:

- <https://jdvelasq.github.io/courses/notebooks/pyspark/1-01-pyspark-wordcount.html>
- <https://jdvelasq.github.io/courses/notebooks/pyspark/1-02-pyspark-app.html>
- <https://jdvelasq.github.io/courses/notebooks/pyspark/1-03-pyspark-operaciones-RDD.html>
- <https://jdvelasq.github.io/courses/notebooks/pyspark/1-04-pyspark-standalone.html>
- <https://jdvelasq.github.io/courses/notebooks/pyspark/2-04-pyspark-SparkSQL.html>
- <https://jdvelasq.github.io/courses/notebooks/pyspark/2-05-pyspark-flights.html>
- <https://jdvelasq.github.io/courses/notebooks/pyspark/3-06-pyspark-classification.html>
- <https://jdvelasq.github.io/courses/notebooks/pyspark/4-07-pyspark-clustering.html>
- <https://jdvelasq.github.io/courses/notebooks/pyspark/5-08-pyspark-regression.html>
- <https://jdvelasq.github.io/courses/notebooks/pyspark/6-09-pyspark-strucStream.html>

ejemplo:

```
Clase_PySpark_WordCount.i Python 3

[1]: ## WordCount en PySpark

[2]: ## Definición del problema
## Se desea contar la frecuencia de las palabras que aparecen en varios archivos de textos. Para simplificar el problema, pruebe el algoritmo con los archivos generados en
## las siguientes celdas.
## Nota.- Los archivos se encuentran en el directorio wordcount de la carpeta de trabajo.

[3]: ##
## Se crea el directorio wordcount en la carpeta actual de trabajo
## y se escriben tres archivos en ella.
##
mkdir -p wordcount/

[4]: %writefile wordcount/text0.txt
Analytics is the discovery, interpretation, and communication of meaningful patterns
in data. Especially valuable in areas rich with recorded information, analytics relies
on the simultaneous application of statistics, computer programming and operations research
to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business
performance. Specifically, areas within analytics include predictive analytics, prescriptive
analytics, enterprise decision management, descriptive analytics, cognitive analytics, Big
Data Analytics, retail analytics, store assortment and stock-keeping unit optimization,
marketing optimization and marketing mix modeling, web analytics, call analytics, speech
analytics, sales force sizing and optimization, price and promotion modeling, predictive
science, credit risk analysis, and fraud analytics. Since analytics can require extensive
computation (see big data), the algorithms and software used for analytics harness the most
current methods in computer science, statistics, and mathematics.

Writing wordcount/text0.txt

[5]: %writefile wordcount/text1.txt
The field of data analysis. Analytics often involves studying past historical data to
research potential trends, to analyze the effects of certain decisions or events, or to
evaluate the performance of a given tool or scenario. The goal of analytics is to improve
the business by gaining knowledge which can be used to make improvements or changes.

Writing wordcount/text1.txt

[6]: %writefile wordcount/text2.txt
Data analytics (DA) is the process of examining data sets in order to draw conclusions
about the information they contain, increasingly with the aid of specialized systems
and software. Data analytics technologies and techniques are widely used in commercial
industries to enable organizations to make more-informed business decisions and by
scientists and researchers to verify or disprove scientific models, theories and
hypotheses.

Writing wordcount/text2.txt
```

En caso de preferir por consola puedes utilizar el editor `pico`.

Los códigos están en [github/AnaliticaGVD](https://github.com/AnaliticaGVD) con sus respectivos resultados de ejecución en la estructura de carpetas por nombre:

amartinUnal / AnaliticaGVD Unwatch 1 Star 0 Fork 0

[Code](#) [Issues 0](#) [Pull requests 0](#) [Actions](#) [Projects 0](#) [Wiki](#) [Security](#) [Insights](#) [Settings](#)

Branch: master **AnaliticaGVD / CodigosClase / 4-Spark /** Create new file Upload files Find file History

amartinUnal summitCodes		Latest commit fa06e1f 26 seconds ago
..		
01-pyspark-wordcount	summitCodes	26 seconds ago
02-pyspark-app	summitCodes	26 seconds ago
03-pyspark-operaciones-RDD	summitCodes	26 seconds ago
04-pyspark-SparkSQL	summitCodes	26 seconds ago
05-pyspark-flights	summitCodes	26 seconds ago
06-pyspark-classification	summitCodes	26 seconds ago
07-pyspark-clustering	summitCodes	26 seconds ago
08-pyspark-regression	summitCodes	26 seconds ago

## Archivo de conteo de palabras:

[amartinUnal](#) / [AnaliticaGVD](#) Unwatch 1 Star 0 Fork 0

[Code](#) [Issues 0](#) [Pull requests 0](#) [Actions](#) [Projects 0](#) [Wiki](#) [Security](#) [Insights](#) [Settings](#)

Branch: master [AnaliticaGVD / CodigosClase / 4-Spark / 01-pyspark-wordcount / 04\\_01\\_Clase\\_PySpark\\_WordCount.ipynb](#) Find file Copy path

[amartinUnal](#) [summitCodes](#) fa06e1f 1 minute ago

1 contributor

675 lines (675 sloc) | 17.7 KB Code Raw Blame History Open Edit Delete

```
In [1]: ## WordCount en PySpark

In [2]: ## Definición del problema
## Se desea contar la frecuencia de las palabras que aparecen en varios archivos de textos. Para simplifi
## car el problema, pruebe el algoritmo con los archivos generados en
## las siguientes celdas.
## Nota.- Los archivos se encuentran en el directorio wordcount de la carpeta de trabajo.

In [3]: ##
## Se crea el directorio wordcount en la carpeta actual de trabajo
## y se escriben tres archivos en ella.
##
!mkdir -p wordcount/

In [4]: %%writefile wordcount/text0.txt
Analytics is the discovery, interpretation, and communication of meaningful patterns
in data. Especially valuable in areas rich with recorded information, analytics relies
on the simultaneous application of statistics, computer programming and operations research
to quantify performance.

Organizations may apply analytics to business data to describe, predict, and improve business
performance. Specifically, areas within analytics include predictive analytics, prescriptive
analytics, enterprise decision management, descriptive analytics, cognitive analytics, Big
Data Analytics, retail analytics, store assortment and stock-keeping unit optimization,
marketing optimization and marketing mix modeling, web analytics, call analytics, speech
analytics, sales force sizing and optimization, price and promotion modeling, predictive
science, credit risk analysis, and fraud analytics. Since analytics can require extensive
```