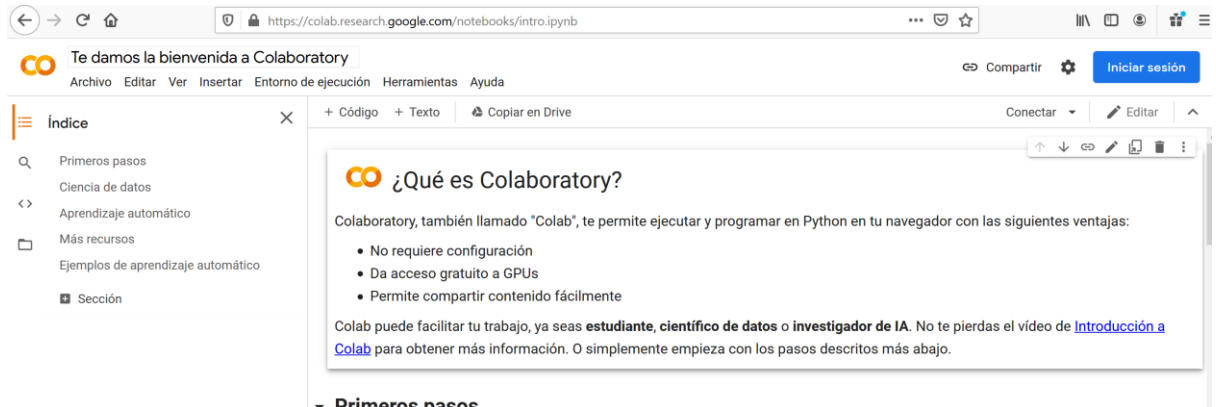


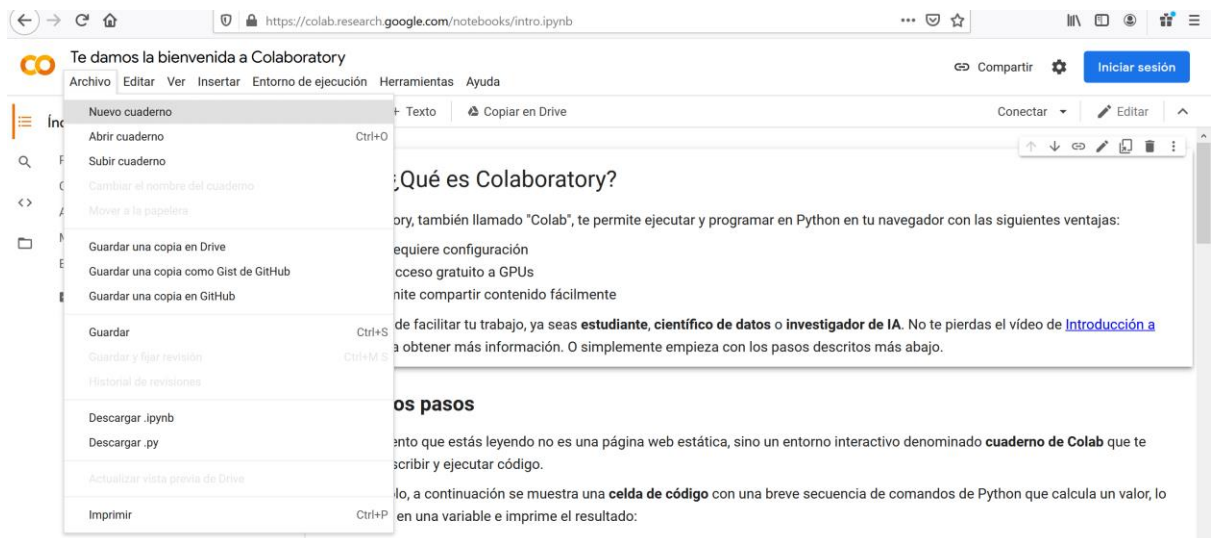
Instalar Hadoop en Google Colab

Google Colab permite crear un entorno de ejecución en la nube de Google de forma gratuita, siempre que Google tenga recursos para ello. Se basa en el notebook Jupyter.

1. Entramos en la página de Google Colab: <https://colab.research.google.com/>



2. Entramos en Archivo -> Nuevo cuaderno:



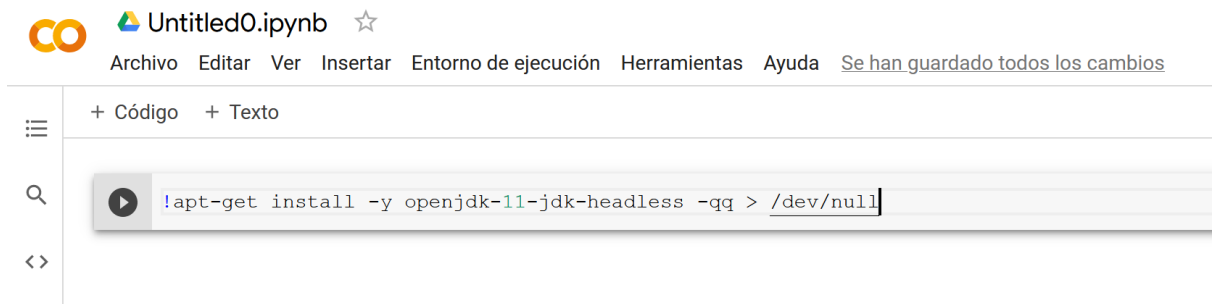
3. Nos dice que tenemos que iniciar sesión con una cuenta de Google. Entonces iniciamos sesión:



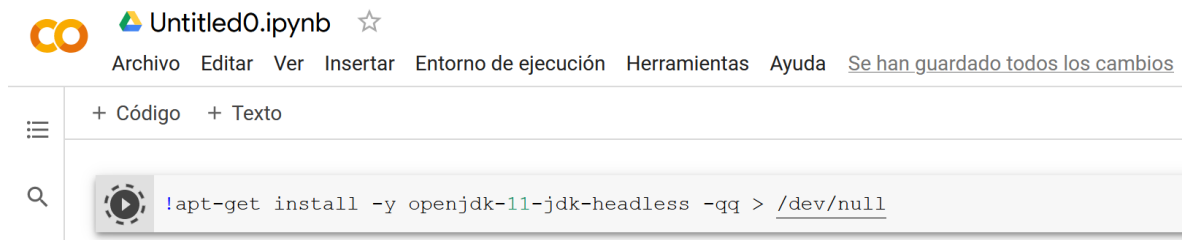
Ya podemos empezar a escribir párrafos (celdas) con código Python, comandos, etc. Para ello tenemos que indicar cuál es el kernel sobre el que se ejecutará la nota

4. Instalamos java, para ello tenemos que indicar: (1) que queremos ejecutar un comando, y (2) que el comando es el de instalar. Es decir, escribimos:

```
!apt-get install -y openjdk-11-jdk-headless -qq > /dev/null
```



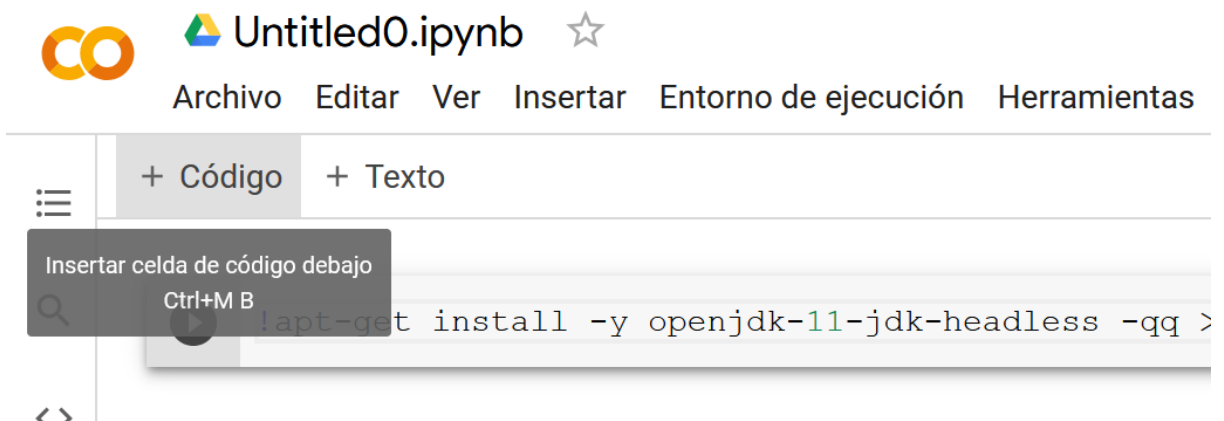
4.1. Le damos a ejecutar a la anterior nota. Para ello pulsamos el botón Play:



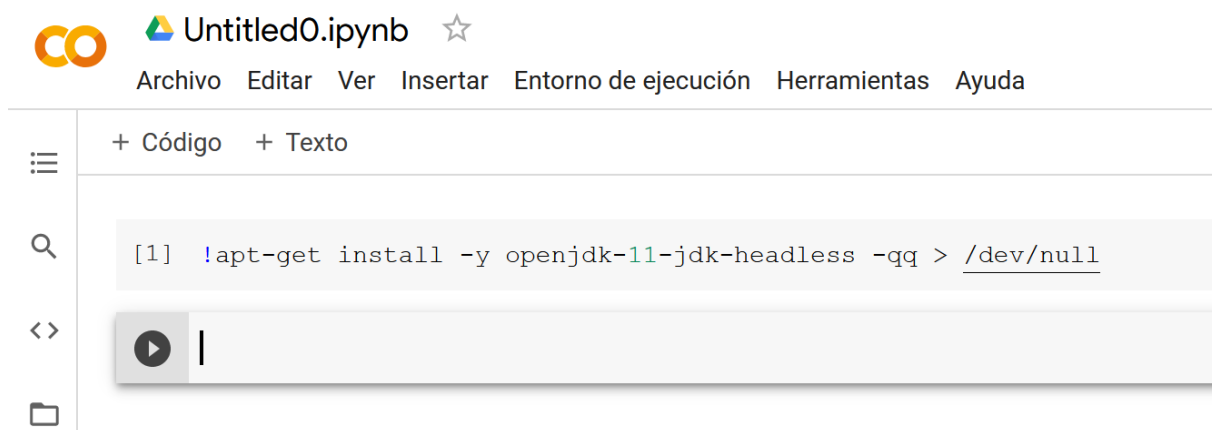
Mientras salga el anterior símbolo, significa que está ejecutando la nota. Puede que tarde unos minutos en instalar java

5. Instalamos Hadoop:

5.1. Creamos una nueva nota, para ello pulsamos en el botón “+ Código”:



Tendremos:



5.2. Descargamos Hadoop: para ello tenemos que indicar:

```
!wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
```



Untitled0.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

+ Código + Texto

```
[1] !apt-get install -y openjdk-11-jdk-headless -qq > /dev/null
```

```
!wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
```

5.3. Pulsamos el botón ejecutar celda (el del símbolo de Play):



Untitled0.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

+ Código + Texto

```
[1] !apt-get install -y openjdk-11-jdk-headless -qq > /dev/null
```

```
!wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
```

```
--2020-12-23 12:58:05-- https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 2a01:4f8:10a:201a::2
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 500749234 (478M) [application/x-gzip]
Saving to: 'hadoop-3.3.0.tar.gz'
```

```
hadoop-3.3.0.tar.gz 16%[==>] 79.34M 49.5MB/s
```

Puede que tarde unos minutos en descargarlo. Pero finalmente tendremos:



Untitled0.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Guardando...

+ Código + Texto

```
[1] !apt-get install -y openjdk-11-jdk-headless -qq > /dev/null
```

```
!wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
```

```
--2020-12-23 12:58:05-- https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 2a01:4f8:10a:201a::2
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 500749234 (478M) [application/x-gzip]
Saving to: 'hadoop-3.3.0.tar.gz'
```

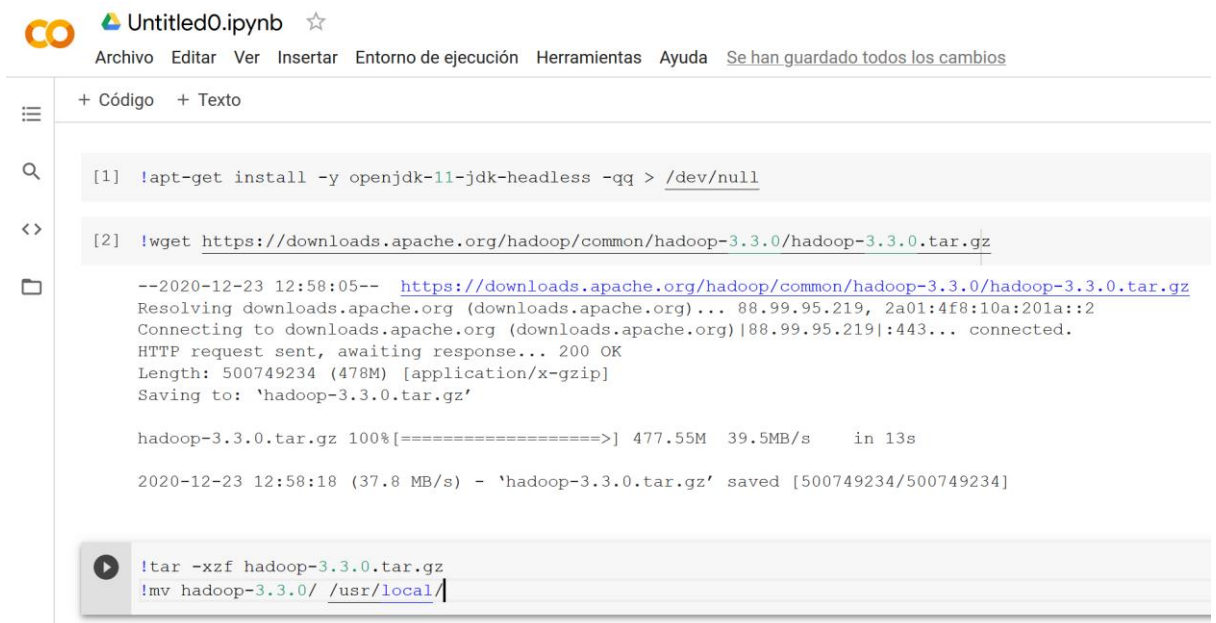
```
hadoop-3.3.0.tar.gz 100%[=====>] 477.55M 39.5MB/s in 13s
```

```
2020-12-23 12:58:18 (37.8 MB/s) - 'hadoop-3.3.0.tar.gz' saved [500749234/500749234]
```

5.4. Ahora añadimos una nueva celda para indicar que descomprima Hadoop y lo mueva a /usr/local. Para ello:

```
!tar -xzf hadoop-3.3.0.tar.gz
```

```
!mv hadoop-3.3.0/ /usr/local/
```



```
[1] !apt-get install -y openjdk-11-jdk-headless -qq > /dev/null

[2] !wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz

--2020-12-23 12:58:05-- https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 2a01:4f8:10a:201a::2
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 500749234 (478M) [application/x-gzip]
Saving to: 'hadoop-3.3.0.tar.gz'

hadoop-3.3.0.tar.gz 100%[=====>] 477.55M 39.5MB/s in 13s

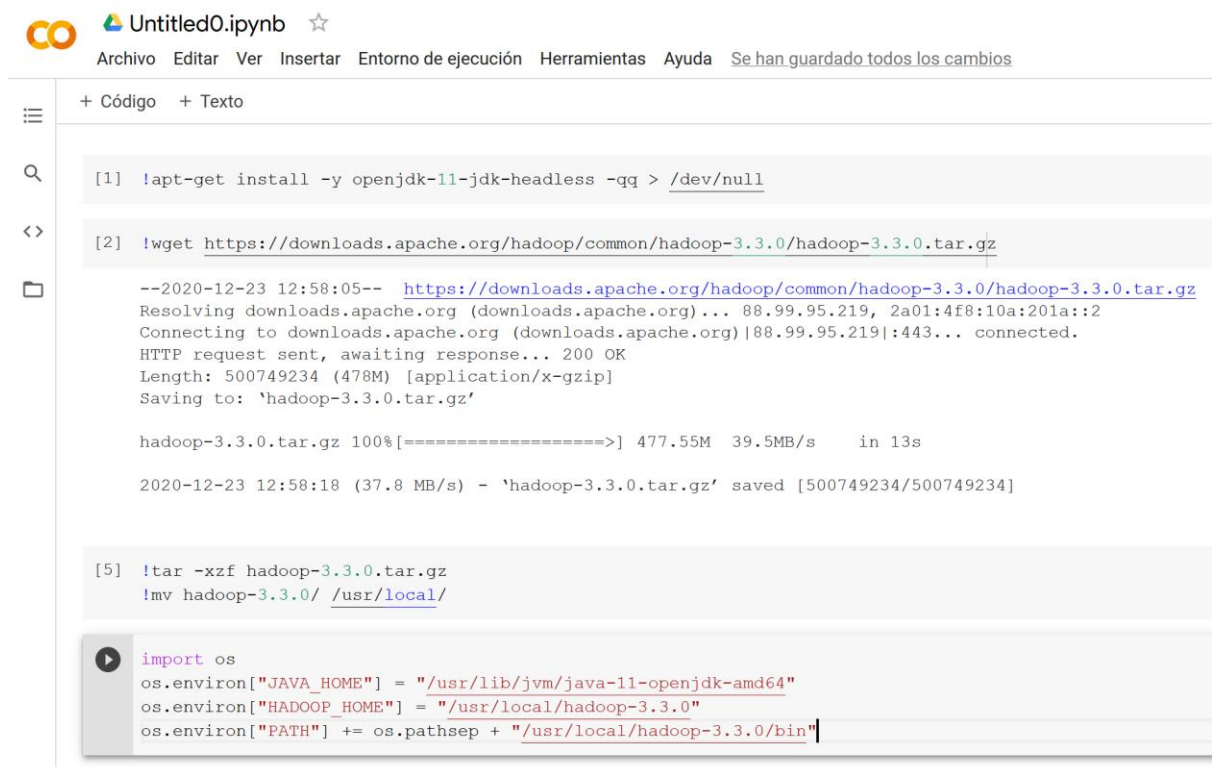
2020-12-23 12:58:18 (37.8 MB/s) - 'hadoop-3.3.0.tar.gz' saved [500749234/500749234]

!tar -xzf hadoop-3.3.0.tar.gz
!mv hadoop-3.3.0/ /usr/local/
```

5.5. Le damos a ejecutar a la última nota que acabamos de crear. Puede que tarde unos minutos en descomprimir.

6. Creamos las variables de entorno de java y hadoop. Para ello creamos una nueva nota con lo siguiente:

```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["HADOOP_HOME"] = "/usr/local/hadoop-3.3.0"
os.environ["PATH"] += os.pathsep + "/usr/local/hadoop-3.3.0/bin"
```



```
[1] !apt-get install -y openjdk-11-jdk-headless -qq > /dev/null

[2] !wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz

--2020-12-23 12:58:05-- https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 2a01:4f8:10a:201a::2
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 500749234 (478M) [application/x-gzip]
Saving to: 'hadoop-3.3.0.tar.gz'

hadoop-3.3.0.tar.gz 100%[=====>] 477.55M 39.5MB/s in 13s

2020-12-23 12:58:18 (37.8 MB/s) - 'hadoop-3.3.0.tar.gz' saved [500749234/500749234]

[3] !tar -xzf hadoop-3.3.0.tar.gz
[4] !mv hadoop-3.3.0/ /usr/local/

[5] import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["HADOOP_HOME"] = "/usr/local/hadoop-3.3.0"
os.environ["PATH"] += os.pathsep + "/usr/local/hadoop-3.3.0/bin"
```

6.1. Le damos a ejecutar la nota

7. Ejecutar programa wordcount:

7.1. Creamos un archivo de entrada, para ello:

```
%%bash
mkdir entradaWordCount
{
    echo "Esto es una linea de prueba"
    echo "segunda linea de prueba"
    echo "Podemos incluir las líneas que queramos"
    echo "esta es la ultima linea"
}> ./entradaWordCount/entrada-1
```



Untitled0.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda



+ Código + Texto



```
[5] !mv hadoop-3.3.0/ /usr/local/
```

```
[19] import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["HADOOP_HOME"] = "/usr/local/hadoop-3.3.0"
os.environ["PATH"] += os.pathsep + "/usr/local/hadoop-3.3.0/bin"
```



```
%%bash
mkdir entradaWordCount
{
    echo "Esto es una linea de prueba"
    echo "segunda linea de prueba"
    echo "Podemos incluir las líneas que queramos"
    echo "esta es la ultima linea"
}> ./entradaWordCount/entrada-1
```

7.1.1. Ejecutamos la celda

7.2. Ejecutamos el programa. Para ello creamos una celda con lo siguiente:

```
!hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar wordcount ./entradaWordCount ./salidaWordCount
```

Untitled0.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Se han guardado todos los cambios

+ Código + Texto

[5] `!mv hadoop-3.3.0/ /usr/local/`

[19] `import os`
`os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"`
`os.environ["HADOOP_HOME"] = "/usr/local/hadoop-3.3.0"`
`os.environ["PATH"] += os.pathsep + "/usr/local/hadoop-3.3.0/bin"`

[47] `%%bash`
`mkdir entradaWordCount`
`{`
 `echo "Esto es una linea de prueba"`
 `echo "segunda linea de prueba"`
 `echo "Podemos incluir las lineas que queramos"`
 `echo "esta es la ultima linea"`
`} > ./entradaWordCount/entrada-1`

`!hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar wordcount ./entradaWordCount ./salidaWordCount`

7.2.1. Ejecutamos la celda

Untitled0.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Se han guardado todos los cambios

+ Código + Texto

[47] `%%bash`
`mkdir entradaWordCount`
`{`
 `echo "Esto es una linea de prueba"`
 `echo "segunda linea de prueba"`
 `echo "Podemos incluir las lineas que queramos"`
 `echo "esta es la ultima linea"`
`} > ./entradaWordCount/entrada-1`

`!hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar wordcount ./entradaWordCount ./salidaWordCount`

2020-12-23 13:49:34,143 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-12-23 13:49:34,446 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-12-23 13:49:34,447 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-12-23 13:49:34,714 INFO input.FileInputFormat: Total input files to process : 1
2020-12-23 13:49:34,751 INFO mapreduce.JobSubmitter: number of splits:1
2020-12-23 13:49:35,176 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2099313454_0001

7.3. Comprobamos la salida, para ello creamos una nota con lo siguiente:

`!cat ./salidaWordCount/*`



+ Código + Texto



```
[48]          Failed Shuffles=0
          Merged Map outputs=1
          GC time elapsed (ms)=0
          Total committed heap usage (bytes)=704643072
          Shuffle Errors
          BAD_ID=0
          CONNECTION=0
          IO_ERROR=0
          WRONG_LENGTH=0
          WRONG_MAP=0
          WRONG_REDUCE=0
          File Input Format Counters
            Bytes Read=117
          File Output Format Counters
            Bytes Written=136
```

<



!cat ./salidaWordCount/*|

7.3.1. Ejecutamos la nota:



+ Código + Texto



```
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=704643072

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=117
File Output Format Counters
  Bytes Written=136
```






```
!cat ./salidaWordCount/*
```

```
Esto      1
Podemos  1
de        2
es        2
esta      1
incluir   1
la        1
las       1
linea     3
líneas    1
prueba    2
que       1
queramos          1
segunda  1
ultima    1
una       1
```





Parte 2: Ejecutar un programa Python en Hadoop

8. Subimos los programas Mapper y Reducer a collab:

8.1. Clickeamos en la parte izquierda de archivos:

  **Untitled0.ipynb** 




Archivo Editar Ver Insertar Entorno de ejecución

+ Código + Texto





```
[48]          Failed Shuffles
              Merged Map output
              GC time elapsed
              Total committed
              Shuffle Errors
              BAD_ID=0
              CONNECTION=0
              IO_ERROR=0
              WRONG_LENGTH=0
              WRONG_MAP=0
              WRONG_REDUCE=0
              File Input Format Count
```



Tendremos:

  **Untitled0.ipynb** 

Archivo Editar Ver Insertar Entorno de ejecución Help

Archivos

..

▶ entradaWordCount

▶ salidaWordCount

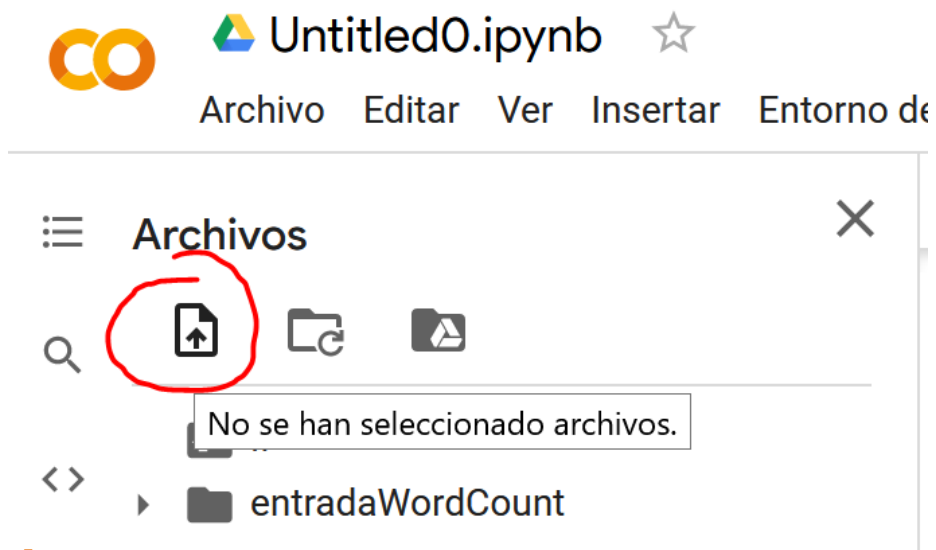
▶ sample_data

hadoop-3.3.0.tar.gz

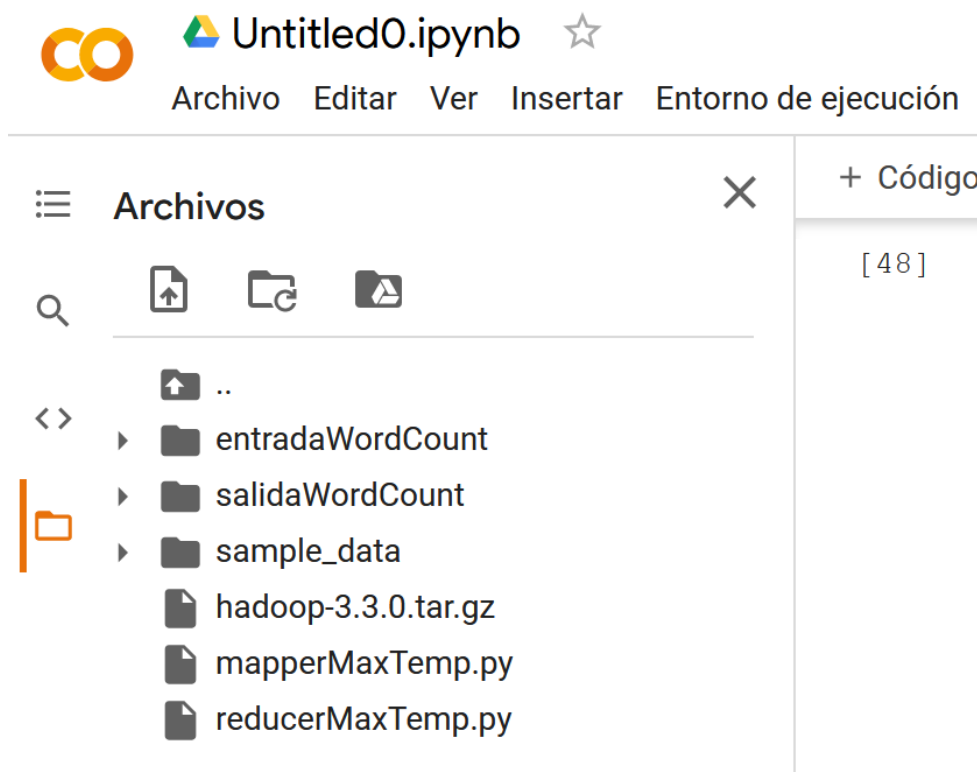
+ Código -

```
[48]
```

8.2. Pulsamos en el botón subir:



8.3. Seleccionamos el archivo Mapper y Reducer que queremos ejecutar. Tendremos:



Importante: Los datos pueden desaparecer cuando reiniciemos google colab. Entonces es recomendable hacer copias de seguridad de los programas por si se pierden

9. Creamos un archivo de entrada para el programa. Para ello creamos la siguiente celda:

```
%%bash
{
    echo -e "1999\tEnero\t1"
    echo -e "1999\tEnero\t5"
    echo -e "1999\tEnero\t3"
    echo -e "1999\tEnero\t2"
    echo -e "1999\tFebrero\t4"
```

```

echo -e "1999\tFebrero\t2"
echo -e "2000\tEnero\t3"
echo -e "2000\tEnero\t6"
echo -e "2000\tFebrero\t6"
echo -e "2000\tFebrero\t2"
echo -e "2001\tAbril\t3"
} > ./medidas.txt

```

The screenshot shows a Jupyter Notebook interface with the following components:

- File Explorer (Archivos):** Displays a directory structure with folders like `entradaWordCount`, `salidaWordCount`, and `sample_data`, and files like `hadoop-3.3.0.tar.gz`, `mapperMaxTemp.py`, and `reducerMaxTemp.py`.
- Code Editor:** Contains a code cell with the command `!cat ./salidaWordCount/*`.
- Terminal Output:** Shows the output of the `cat` command, displaying word counts for various words.

Esto	1
Podemos	1
de	2
es	2
esta	1
incluir	1
la	1
las	1
línea	3
líneas	1
prueba	2
que	1
queramos	1
segunda	1
ultima	1
una	1
- Terminal:** Shows a bash shell session with a code block containing the same `echo` commands as the first block, followed by `> ./medidas.txt`.

9.1. Ejecutamos la nota

10. Ejecutamos el programa:

- 10.1. Le damos permisos de ejecución a los programas Mapper y Reducer. Para ello creamos la siguiente celda:

```
!chmod u+x ./mapperMaxTemp.py
```

```
!chmod u+x ./reducerMaxTemp.py
```

Untitled0.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Se han guardado todos los cambios

Archivos

- ..
- entradaWordCount
- salidaWordCount
- sample_data
- hadoop-3.3.0.tar.gz
- mapperMaxTemp.py
- medidas.txt
- reducerMaxTemp.py

+ Código + Texto

```
[51] %%bash
{
  echo -e "1999\tEnero\t1"
  echo -e "1999\tEnero\t5"
  echo -e "1999\tEnero\t3"
  echo -e "1999\tEnero\t2"
  echo -e "1999\tFebrero\t4"
  echo -e "1999\tFebrero\t2"
  echo -e "2000\tEnero\t3"
  echo -e "2000\tEnero\t6"
  echo -e "2000\tFebrero\t6"
  echo -e "2000\tFebrero\t2"
  echo -e "2001\tAbril\t3"
} > ./medidas.txt
```

```
!chmod u+x ./mapperMaxTemp.py
!chmod u+x ./reducerMaxTemp.py
```

10.1.1. Ejecutamos la celda

10.2. Ejecutamos el programa. Para ello creamos una celda con el siguiente comando:

```
!hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.0.jar -
file ./mapperMaxTemp.py -mapper ./mapperMaxTemp.py -file
./reducerMaxTemp.py -reducer ./reducerMaxTemp.py -input medidas.txt -output
./salidaMaxTemp1
```

Untitled0.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Se han guardado todos los cambios

Comentario Compartir Editar

Archivos

- ..
- entradaWordCount
- salidaWordCount
- sample_data
- hadoop-3.3.0.tar.gz
- mapperMaxTemp.py
- medidas.txt
- reducerMaxTemp.py

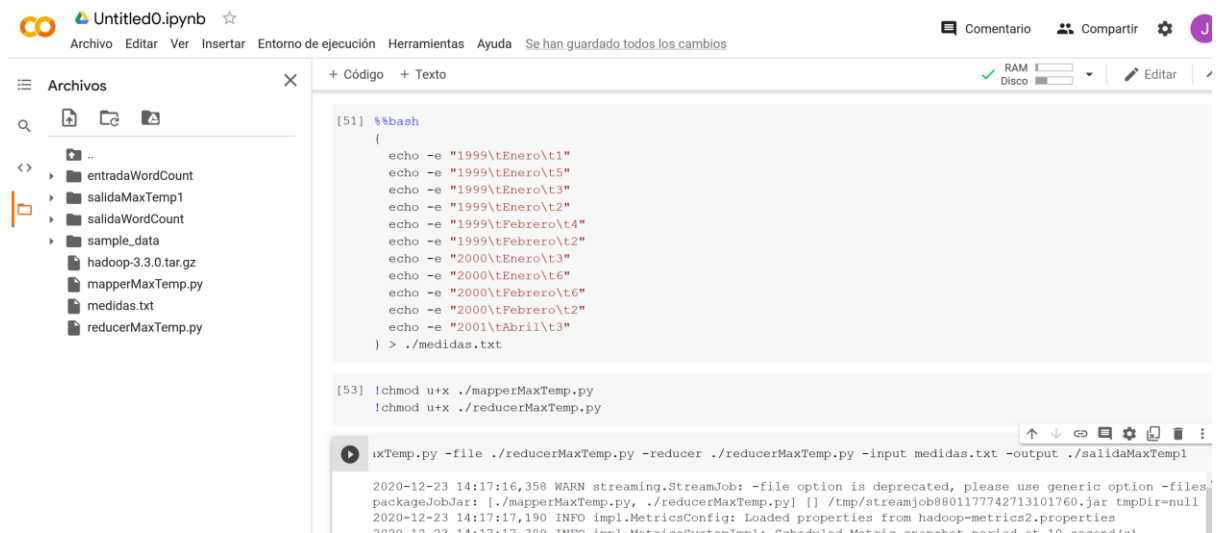
+ Código + Texto

```
[51] %%bash
{
  echo -e "1999\tEnero\t1"
  echo -e "1999\tEnero\t5"
  echo -e "1999\tEnero\t3"
  echo -e "1999\tEnero\t2"
  echo -e "1999\tFebrero\t4"
  echo -e "1999\tFebrero\t2"
  echo -e "2000\tEnero\t3"
  echo -e "2000\tEnero\t6"
  echo -e "2000\tFebrero\t6"
  echo -e "2000\tFebrero\t2"
  echo -e "2001\tAbril\t3"
} > ./medidas.txt
```

```
[53] !chmod u+x ./mapperMaxTemp.py
!chmod u+x ./reducerMaxTemp.py
```

```
ixTemp.py -file ./reducerMaxTemp.py -reducer ./reducerMaxTemp.py -input medidas.txt -output ./salidaMaxTemp1
```

10.2.1. Ejecutamos la celda. Tendremos:



The screenshot shows a Jupyter Notebook titled 'Untitled0.ipynb'. The left sidebar displays a file explorer with the following structure:

- ..
- entradaWordCount
- salidaMaxTemp1
- salidaWordCount
- sample_data
- hadoop-3.3.0.tar.gz
- mapperMaxTemp.py
- medidas.txt
- reducerMaxTemp.py

The main code area contains the following commands:

```
[51] %%bash
{
  echo -e "1999\tEnero\t1"
  echo -e "1999\tEnero\t5"
  echo -e "1999\tEnero\t3"
  echo -e "1999\tEnero\t2"
  echo -e "1999\tFebrero\t4"
  echo -e "1999\tFebrero\t2"
  echo -e "2000\tEnero\t3"
  echo -e "2000\tEnero\t6"
  echo -e "2000\tFebrero\t6"
  echo -e "2000\tFebrero\t2"
  echo -e "2001\tAbril\t3"
} > ./medidas.txt

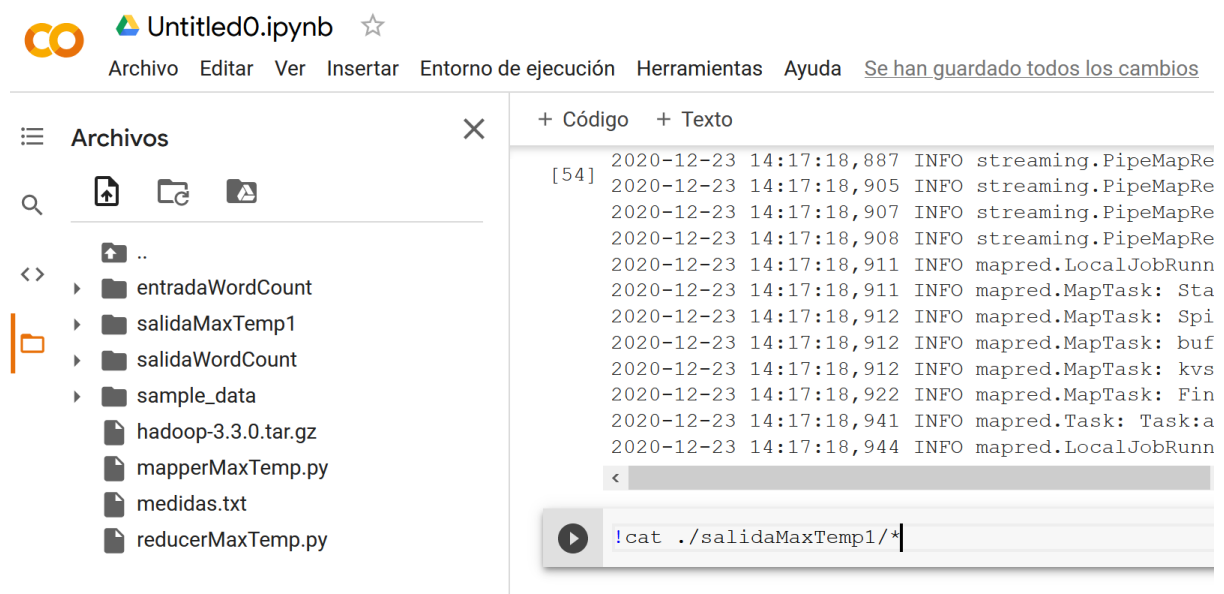
[53] !chmod u+x ./mapperMaxTemp.py
!chmod u+x ./reducerMaxTemp.py

!xTemp.py -file ./reducerMaxTemp.py -reducer ./reducerMaxTemp.py -input medidas.txt -output ./salidaMaxTemp1
```

The output of the last command shows a warning from streaming.StreamJob and a successful execution of the MapReduce job.

11. Comprobamos la salida. Para ello creamos la siguiente celda:

`!cat ./salidaMaxTemp1/*`



The screenshot shows a Jupyter Notebook titled 'Untitled0.ipynb'. The left sidebar displays the same file explorer as the previous screenshot.

The main code area contains the following command:

```
[54] !cat ./salidaMaxTemp1/*
```

The output of the command shows the contents of the 'salidaMaxTemp1' directory, which contains a single file named 'salidaMaxTemp1.txt'. The file contains the following text:

```
2020-12-23 14:17:16,358 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files
packageJobJar: [./mapperMaxTemp.py, ./reducerMaxTemp.py] [] /tmp/streamjob880117742713101760.jar tmpDir=null
2020-12-23 14:17:17,190 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-12-23 14:17:17,200 INFO impl.MetricsReporter: Scheduled Metrics reporter period at 10 second(s)
```

11.1. Ejecutamos la celda:



Archivos



..

▶ entradaWordCount

▶ salidaMaxTemp1

▶ salidaWordCount

▶ sample_data

hadoop-3.3.0.tar.gz

mapperMaxTemp.py

medidas.txt

reducerMaxTemp.py

+ Código + Texto

```
[54] 2020-12-23 14:17:18,887 INFO streaming.PipeMapRed:
2020-12-23 14:17:18,905 INFO streaming.PipeMapRed:
2020-12-23 14:17:18,907 INFO streaming.PipeMapRed:
2020-12-23 14:17:18,908 INFO streaming.PipeMapRed:
2020-12-23 14:17:18,911 INFO mapred.LocalJobRunner
2020-12-23 14:17:18,911 INFO mapred.MapTask: Start
2020-12-23 14:17:18,912 INFO mapred.MapTask: Spill
2020-12-23 14:17:18,912 INFO mapred.MapTask: bufst
2020-12-23 14:17:18,912 INFO mapred.MapTask: kvsta
2020-12-23 14:17:18,922 INFO mapred.MapTask: Finis
2020-12-23 14:17:18,941 INFO mapred.Task: Task:att
2020-12-23 14:17:18,944 INFO mapred.LocalJobRunner
```



```
!cat ./salidaMaxTemp1/*
```

```
1999    5.0
2000    6.0
2001    3.0
```