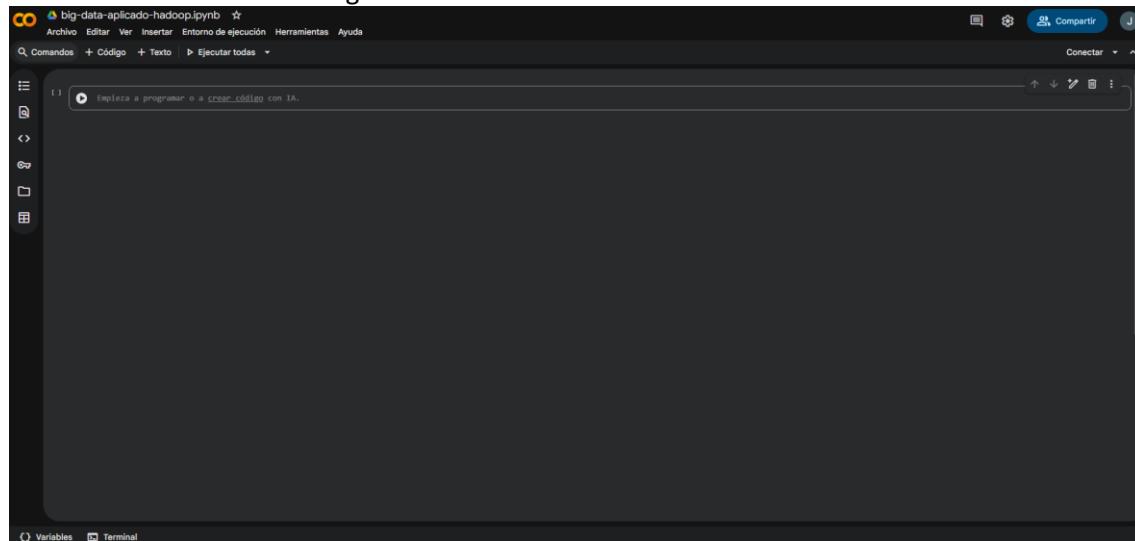


Solución tarea 2.1

1. Creo un notebook en Google Colab



2. Instalo hadoop: Hemos descargado la versión 3.4.2 que es la más reciente

```
[1] wget https://archive.apache.org/dist/hadoop/common/hadoop-3.4.2/hadoop-3.4.2.tar.gz
...
[2] !tar -xzf hadoop-3.4.2.tar.gz
```

3. Descomprimimos y movemos la carpeta al lugar que nos indican

```
[2] !tar -xzf hadoop-3.4.2.tar.gz
[3] !mv hadoop-3.4.2/ /usr/local/
```

4. Configuraremos el Hadoop JAVA HOME

- Buscamos la dirección de Java en la máquina Google Colab y establecemos el valor de la variable

```
Archivos          EJ
+-- Archivos
    +-- hadoop-3.4.2
    +-- sample_data
    +-- hadoop-3.4.2.tar.gz

[1] !readlink -f /usr/bin/java | sed "s/bin/java:/"
[1] /usr/lib/jvm/java-17-openjdk-amd64/
[2] !import os
[2] os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-17-openjdk-amd64/"
```

5. Ejecutamos Hadoop - Comandos de prueba

```
Archivos          (9) ✓ 1s
          ↗ !/usr/local/hadoop-3.4.2/bin/hadoop version
          ...
          * Hadoop 3.4.2
          * Source code repository https://github.com/apache/hadoop.git - r-84e8b89ee2ebef69236f1205be171badde7a495c
          * Compiled by ambrusw on 2025-08-20T10:36Z
          * Compiled on platform linux-x86_64
          * Configuration from /etc/hadoop/hadoop-env.sh
          * From source with checksum f494c67d4b4be21bfbe9515c9b0f7f6
          This command was run using /usr/local/hadoop-3.4.2/share/hadoop/common/hadoop-common-3.4.2.jar
```

- Jar de ejemplos

```
[8]  !mkdir ~/input
```

[9]  ✓ 0 s !cp /usr/local/hadoop-3.4.2/etc/hadoop/*.xml ~/input

Ejemplo 1: Comando Grep que sirve para buscar patrones de texto dentro de ficheros

```
[10] ✓ 0s !ls ~/input
...
*** capacity-scheduler.xml hdfs-rbf-site.xml kms-acls.xml     yarn-site.xml
core-site.xml          hdfs-site.xml    kms-site.xml
hadoop-policy.xml      httpfs-site.xml mapred-site.xml

⌚ %bash
rm -rf ~/grep_example
mkdir -p ~/input
echo "this line is allowed" > ~/input/file1.txt

/usr/local/hadoop-3.4.2/bin/hadoop jar \
/usr/local/hadoop-3.4.2/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.2.jar \
grep ~/input ~/grep_example 'allowed[.]'

...
2025-12-16 12:29:23,465 INFO input.FileInputFormat: Total input files to process : 1
2025-12-16 12:29:23,511 INFO mapreduce.JobSubmitter: number of splits:1
2025-12-16 12:29:23,979 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local276438840_0001
2025-12-16 12:29:23,981 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-12-16 12:29:24,304 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-12-16 12:29:24,309 INFO mapreduce.Job: Running job: job_local276438840_0001
2025-12-16 12:29:24,312 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-12-16 12:29:24,327 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to PathOutputCommitterFactory
2025-12-16 12:29:24,328 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-12-16 12:29:24,329 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-12-16 12:29:24,329 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-12-16 12:29:24,411 INFO mapred.LocalJobRunner: Starting task: attempt_local276438840_0001_m_000000_0
2025-12-16 12:29:24,412 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-12-16 12:29:24,450 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to PathOutputCommitterFactory
2025-12-16 12:29:24,450 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-12-16 12:29:24,484 INFO mapred.Task: Using ResourceCalculatorProcessTree : []
2025-12-16 12:29:24,756 INFO mapred.MapTask: Processing split: file:/root/input/file1.txt:0+21
2025-12-16 12:29:24,756 INFO mapred.MapTask: (EQUATOR) 0 kvl 26214396, ¿Cómo puedo instalar bibliotecas de Python? Carga datos desde Google Drive Muéstrame un
2025-12-16 12:29:24,756 INFO mapred.MapTask: soft limit at 83886080
2025-12-16 12:29:24,756 INFO mapred.MapTask: bufstart = 0; bufvoid =
2025-12-16 12:29:24,756 INFO mapred.MapTask: kvstart = 26214396; leng
2025-12-16 12:29:24,763 INFO mapred.MapTask: Map output collector cl
2025-12-16 12:29:24,776 INFO mapred.MapTask: Starting flush of map o
2025-12-16 12:29:24,777 INFO mapred.MapTask: Starting flush of map o
2025-12-16 12:29:24,777 INFO mapred.MapTask: Spilling map output
```

```
[10] ✓ 0s !cat ~/grep_example/*
...
*** 1      allowed
```

Ejemplo 2: WordCount que se utiliza para contar el número de apariciones de cada palabra en un conjunto de archivos

```
[19] [ 0 s] !head ~/wordcount_output/part-r-00000
... allowed 1
      is      1
      line    1
      this   1
```

Ejemplo 3: Sort para ordenar grandes volúmenes de datos.

```
[25] ✓ 12.6
  ➜ %%bash
    /usr/local/hadoop-3.4.2/bin/hadoop jar \
    /usr/local/hadoop-3.4.2/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.2.jar \
    sort ~/sort_input ~/sort_output

2025-12-16 12:58:15,730 INFO output.FileOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-12-16 12:58:15,730 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-12-16 12:58:15,730 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-12-16 12:58:15,730 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-12-16 12:58:15,730 INFO mapred.MapTask: Processing split: file:/root/sort_input/part-m-00000:872415232+33554432
2025-12-16 12:58:15,893 INFO mapred.LocalJobRunner:
2025-12-16 12:58:15,893 INFO mapred.Task: Task attempt_local1154341909_0001_m_000027_0 is done. And is in the process of committing
2025-12-16 12:58:15,893 INFO mapred.LocalJobRunner:
2025-12-16 12:58:15,893 INFO mapred.Task: Task attempt_local1154341909_0001_m_000027_0 is allowed to commit now
2025-12-16 12:58:15,893 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1154341909_0001_m_000027_0' to file:/root/sort_output
2025-12-16 12:58:15,895 INFO mapred.LocalJobRunner: map
2025-12-16 12:58:15,896 INFO mapred.Task: Task 'attempt_local1154341909_0001_m_000027_0' done.
2025-12-16 12:58:15,896 INFO mapred.Task: Final Counters for attempt_local1154341909_0001_m_000027_0: Counters: 15
  System Counters
    FILE: Number of bytes read=951126287
    FILE: Number of bytes written=949685159
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=3177
    Map output records=3177
    Input split bytes=99
    Spilled Records: 1 |Carga datos desde Google Drive |Muéstrame un
    Failed Shuffles=0
    Merged Maps
    GC time elas... |Qué puedo ayudarte a crear?
    Total committed memory=0
  File Input Format: + |>
    Bytes Read=.....
```

Ejemplo 4: Copando Pi para calcular una aproximación del número pi

```
[27] 5s
%%bash
/usr/local/hadoop-3.4.2/bin/hadoop jar \
/usr/local/hadoop-3.4.2/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.4.2.jar \
pi 10 1000

...
*** Number of Maps = 10
Samples per Map = 1000
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Starting Job
Job Finished in 3.242 seconds
Estimated value of Pi is 3.14080000000000000000
2025-12-10 12:59:26,310 INFO InputFileInputFormat: Total input files to process : 10
2025-12-10 12:59:26,330 INFO mapreduce.JobSubmitter: number of splits:10
2025-12-10 12:59:26,678 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1449526893_0001
2025-12-10 12:59:26,688 INFO mapreduce.JobSubmitter: Executing with tokens: []
```