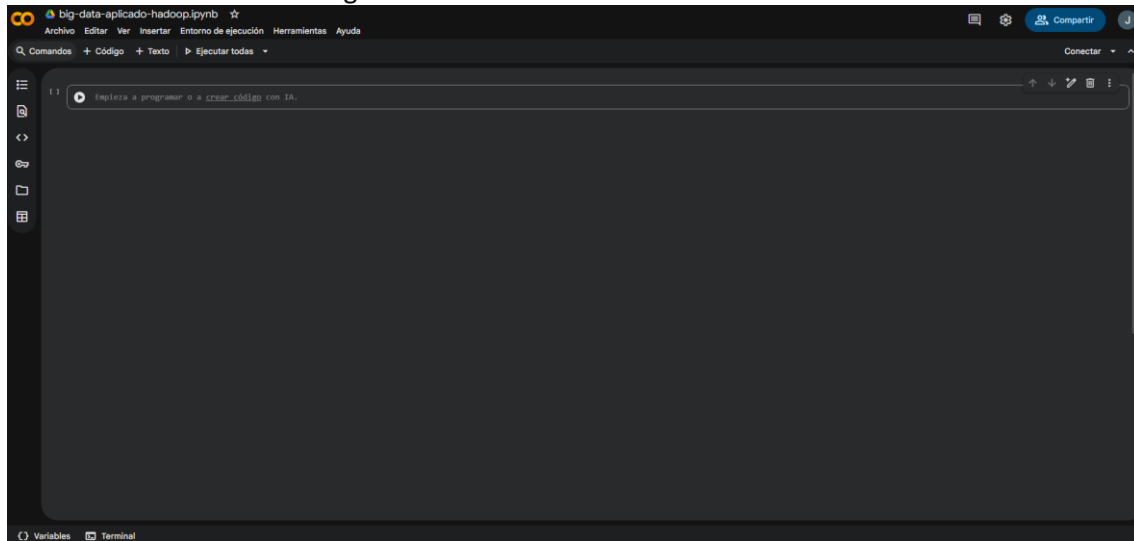


Solución tarea 2.1

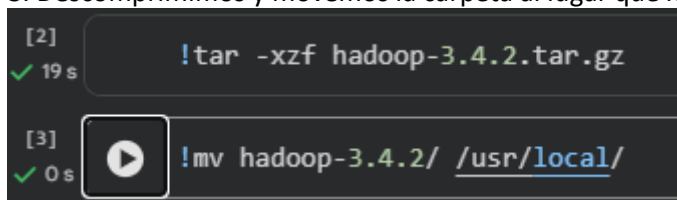
1. Creo un notebook en Google Colab



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



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




4. Configuramos el Hadoop JAVA HOME

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



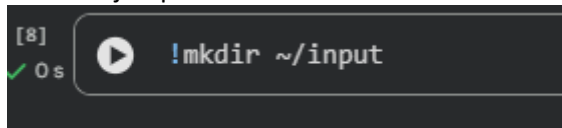
5. Ejecutamos Hadoop

- Comandos de prueba

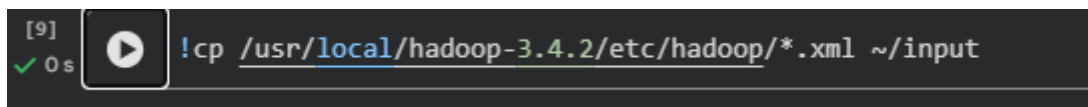


```
Archivos [1] /usr/local/hadoop-3.4.2/bin/hadoop version
--
Hadoop 3.4.2
Source code repository https://github.com/apache/hadoop.git -r 84e8b89e2e6d23691265b9e171bade7a495c
Compiled by almaru on 2025-08-20T18:30Z
Compiled on platform linux_x86_64
Compiled with protoc 3.23.4
From source with checksum fa94c67d4d44ee821b9e9515c9b0f7b6
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
✓ 0s
```



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

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

```
[10] 0s ll 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 12:29:24,308 INFO mapreduce.Job: The url to track the job: http://localhost:8880/
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 FileOutputCommitterFactory
2025-12-16 12:29:24,328 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-12-16 12:29:24,328 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 FileOutputCommitterFactory
2025-12-16 12:29:24,450 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-12-16 12:29:24,450 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-12-16 12:29:24,484 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-12-16 12:29:24,500 INFO mapred.MapTask: Processing split: file:/root/.input/file1.txt:0+21
2025-12-16 12:29:24,756 INFO mapred.MapTask: (EQUATOR) 0 kv1 26214396 ¿Cómo puedo instalar bibliotecas de Python? Carga datos desde Google Drive Muéstrame un
2025-12-16 12:29:24,756 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-12-16 12:29:24,756 INFO mapred.MapTask: soft limit at 838860800
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; lenq
2025-12-16 12:29:24,763 INFO mapred.MapTask: Map output collector cl
2025-12-16 12:29:24,776 INFO mapred.LocalJobRunner:
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

```
[18]   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 \  
wordcount ~/input ~/wordcount_output  
  
2025-12-16 12:54:08,888 INFO mapred.maptask: kvstart = 20214395(10483/584); kvend = 20214394(10483/530); length = 13/0>>>3000  
2025-12-16 12:54:08,897 INFO mapred.MapTask: finished spill 0  
2025-12-16 12:54:08,911 INFO mapred.Task: Task:attempt_local630758942_0001_m_000000_0 is done. And is in the process of committing  
2025-12-16 12:54:08,918 INFO mapred.LocalJobRunner: map  
2025-12-16 12:54:08,918 INFO mapred.Task: Task "attempt_local630758942_0001_m_000000_0" done.  
2025-12-16 12:54:08,931 INFO mapred.Task: Final Counters for attempt_local630758942_0001_m_000000_0: Counters: 18  
File System Counters  
FILE: Number of bytes read=281799  
FILE: Number of bytes written=903692  
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=1  
Map output records=4  
Map output bytes=37  
Map output materialized bytes=51  
Input split bytes=91  
Combine input records=4  
Combine output records=4  
Spilled Records=4  
Failed Shuffles=0  
Merged Map outputs=0  
GC time elapsed (ms)=0  
Total committed heap usage (bytes)=297795584  
File Input Format Counters  
Bytes Read=281799  
2025-12-16 12:54:08,931 INFO mapred.LocalJobRunner: finishing Task: attempt_local630758942_0001_m_000000_0  
2025-12-16 12:54:08,932 INFO mapred.Task: Task:attempt_local630758942_0001_m_000000_0 is done. And is in the process of committing  
2025-12-16 12:54:08,938 INFO mapred.LocalJobRunner: map  
2025-12-16 12:54:08,938 INFO mapred.Task: Task "attempt_local630758942_0001_m_000000_0" done.  
2025-12-16 12:54:08,950 INFO mapred.Task: Final Counters for attempt_local630758942_0001_m_000000_0: Counters: 18  
2025-12-16 12:54:08,950 INFO mapred.Task: Task:attempt_local630758942_0001_m_000000_0 is done. And is in the process of committing  
2025-12-16 12:54:08,959 INFO mapred.LocalJobRunner: map  
2025-12-16 12:54:08,959 INFO mapred.Task: Task "attempt_local630758942_0001_m_000000_0" done.  
2025-12-16 12:54:08,951 INFO mapred.Task: Final Counters for attempt_local630758942_0001_m_000000_0: Counters: 18  
2025-12-16 12:54:08,955 INFO mapred.Task: Task:attempt_local630758942_0001_m_000000_0 is done. And is in the process of committing
```

```
[19]   !head ~/wordcount_output/part-r-000000  
  
... allowed 1  
is 1  
line 1  
this 1
```

Ejemplo 3: Sort para ordenar grandes volúmenes de datos.

```
[25] %bash
13 s
/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 12:58:15,736 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-12-16 12:58:15,736 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-12-16 12:58:15,737 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-12-16 12:58:15,738 INFO mapred.MapTask: Processing split: file:/root/sort_input/part-m-00000:872415232+33554432
2025-12-16 12:58:15,881 INFO mapred.LocalJobRunner:
2025-12-16 12:58:15,892 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
File 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 Rec: ¿Cómo puedo instalar bibliotecas de Python? Carga datos desde Google Drive Muestrame un
  Failed Shuffles=0
  Merged Map
  GC time el: ¿Qué puedo ayudarte a crear?
  Total comm:
  File Input Format ( +
  Bytes Read: 951126287
```

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

```
[27] 5s  Xxhbash
/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 3.140800000000000000000000
2025-12-16 12:59:26,310 INFO input.FileInputFormat: Total input files to process : 10
2025-12-16 12:59:26,330 INFO mapreduce.JobSubmitter: number of splits:10
2025-12-16 12:59:26,678 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1449526893_0001
2025-12-16 12:59:26,680 INFO mapreduce.JobSubmitter: Executing with tokens: []
```