AWS CLI.
Como resultado de la práctica has de entregar un archivo en formato *"pdf"* con breves explicaciones y capturas de pantalla de los pasos principales que has dado para realizar los ejercicios.

CONTENIDO
APARTADO A

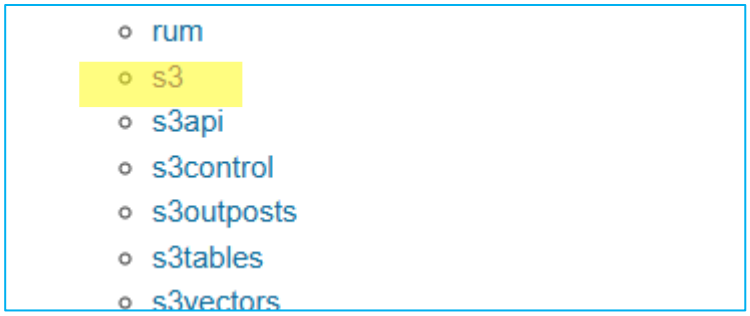
INTRODUCCIÓN

Investiga en Internet que comandos de AWS CLI hemos de utilizar para manejar *buckets* S3.
Por ejemplo, en el siguiente enlace hay información de AWS sobre ello:

https://docs.aws.amazon.com/cli/latest/userguide/cli-services-s3-commands.html

Toda la lista de comando de AWS CLI para S3

s3 — AWS CLI 2.32.32 Command Reference



Utilizando únicamente comando AWS CLI:

1. Crear tres *buckets* en la región por defecto (con nombres similares a: **primero**, **copia**, **sincro**).

Syntax

```
$ aws s3 mb <target> [--options]
```

s3 mb examples

The following example creates the `s3://amzn-s3-demo-bucket` bucket.

```
$ aws s3 mb s3://amzn-s3-demo-bucket
```

Creamos los buckets

primeros3-brayan

copias3-brayan

sincro3-brayan

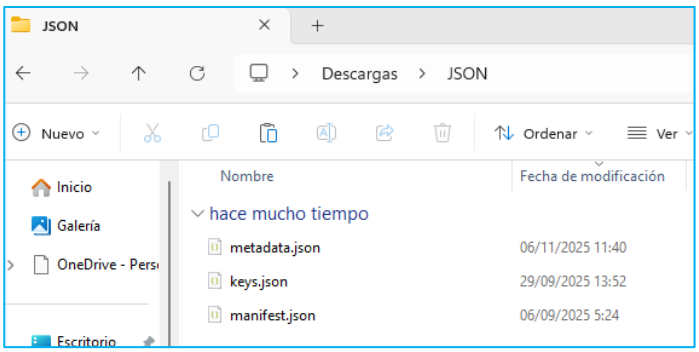
```
PS C:\Users\Mañana> aws s3 mb s3://primeros3-brayan
make_bucket: primeros3-brayan
PS C:\Users\Mañana> aws s3 mb s3://copias3-brayan
make_bucket: copias3-brayan
PS C:\Users\Mañana> aws s3 mb s3://sincro3-brayan
make_bucket: sincro3-brayan
PS C:\Users\Mañana>
```

2. Listar todos los cubos.

```
PS C:\Users\Mañana> aws s3 ls
2026-01-08 08:51:55 aws-logs-204954676159-us-east-1
2026-01-12 09:39:34 copias3-brayan
2026-01-12 09:39:17 primeros3-brayan
2026-01-12 09:39:43 sincro3-brayan
PS C:\Users\Mañana>
```

3. Ponte en una carpeta donde tengas varios archivos, mejor CSV's o JSON, Sube al menos dos archivos de esa carpeta al *bucket* **primero**.

   Tenemos 3 archivo de eclipse



```
PS C:\Users\Mañana> cd C:\Users\Mañana\Downloads\JSON
PS C:\Users\Mañana\Downloads\JSON> ls


    Directorio: C:\Users\Mañana\Downloads\JSON


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----        29/09/2025     13:52           8756 keys.json
-a----        06/09/2025      5:24            113 manifest.json
-a----        06/11/2025     11:40           1092 metadata.json
```

```
PS C:\Users\Mañana\Downloads\JSON> aws s3 cp keys.json s3://primeros3-brayan
upload: .\keys.json to s3://primeros3-brayan/keys.json
PS C:\Users\Mañana\Downloads\JSON> aws s3 cp manifest.json s3://primeros3-brayan
upload: .\manifest.json to s3://primeros3-brayan/manifest.json
PS C:\Users\Mañana\Downloads\JSON>
```

4. Copia los archivos del *bucket* **primero** a **copia** de uno en uno.

```
PS C:\Users\Mañana\Downloads\JSON> aws s3 cp s3://primeros3-brayan/keys.json s3://copias3-brayan
copy: s3://primeros3-brayan/keys.json to s3://copias3-brayan/keys.json
PS C:\Users\Mañana\Downloads\JSON> aws s3 cp s3://primeros3-brayan/manifest.json s3://copias3-brayan
copy: s3://primeros3-brayan/manifest.json to s3://copias3-brayan/manifest.json
PS C:\Users\Mañana\Downloads\JSON>
```

5. Mover todos los objetos de **primero** a **copia** dentro de la carpeta **datos** de forma recursiva.

```
PS C:\Users\Mañana\Downloads\JSON> aws s3 mv s3://primeros3-brayan s3://copias3-brayan/datos --recursive
move: s3://primeros3-brayan/manifest.json to s3://copias3-brayan/datos/manifest.json
move: s3://primeros3-brayan/keys.json to s3://copias3-brayan/datos/keys.json
```

6. Listar los objetos de un **copia/datos**.

```
PS C:\Users\Mañana\Downloads\JSON> aws s3 ls s3://copias3-brayan/datos/
2026-01-12 09:56:42       8756 keys.json
2026-01-12 09:56:42        113 manifest.json
```

7. Listar recursivamente todos los objetos el cubo **copia.**

```
PS C:\Users\Mañana\Downloads\JSON> aws s3 ls s3://copias3-brayan --recursive
2026-01-12 09:56:42       8756 datos/keys.json
2026-01-12 09:56:42        113 datos/manifest.json
```

8. Lista recursivamente los datos del cubo en formato *legible por humanos*. ¿Qué significa eso?

```
PS C:\Users\Mañana\Downloads\JSON> aws s3 ls s3://copias3-brayan --recursive
2026-01-12 09:56:42       8756 datos/keys.json
2026-01-12 09:56:42        113 datos/manifest.json
PS C:\Users\Mañana\Downloads\JSON> aws s3 ls s3://copias3-brayan --recursive --human-readable
2026-01-12 09:56:42    8.6 KiB datos/keys.json
2026-01-12 09:56:42  113 Bytes datos/manifest.json
PS C:\Users\Mañana\Downloads\JSON>
```

Que los datos aparezcan con su unidad de medida en B, Kb, Mb O Gb que son unidades más entendibles que solo muestre los bytes para los humanos

9. Mostrar la información resumida, incluido el número de objetos y el tamaño total del cubo **copia**.

```
PS C:\Users\Mañana\Downloads\JSON> aws s3 ls s3://copias3-brayan --recursive --summarize
2026-01-12 09:56:42       8756 datos/keys.json
2026-01-12 09:56:42        113 datos/manifest.json

Total Objects: 2
   Total Size: 8869
```

10. Muévete a otra carpeta en local y descarga todos los objetos de **copia** a ella. Borra en local los archivos descargados.

```
PS C:\Users\Mañana\Downloads\JSON> mkdir DESCARGA_S3
```

```
PS C:\Users\Mañana\Downloads\JSON\DESCARGA_S3> aws s3 cp s3://copias3-brayan/ . --recursive
download: s3://copias3-brayan/datos/manifest.json to datos\manifest.json
download: s3://copias3-brayan/datos/keys.json to datos\keys.json
PS C:\Users\Mañana\Downloads\JSON\DESCARGA_S3> ls


    Directorio: C:\Users\Mañana\Downloads\JSON\DESCARGA_S3


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
d-----        12/01/2026     10:14                datos


PS C:\Users\Mañana\Downloads\JSON\DESCARGA_S3> rm -r datos
PS C:\Users\Mañana\Downloads\JSON\DESCARGA_S3> ls
PS C:\Users\Mañana\Downloads\JSON\DESCARGA_S3>
```

11. Sincronizar el contenido del un cubo **copia** con el cubo **sincro.**

```
PS C:\Users\Mañana\Downloads\JSON\DESCARGA_S3> aws s3 sync s3://copias3-brayan/ s3://sincro3-brayan
copy: s3://copias3-brayan/datos/keys.json to s3://sincro3-brayan/datos/keys.json
copy: s3://copias3-brayan/datos/manifest.json to s3://sincro3-brayan/datos/manifest.json
```

12. Sincronizar el contenido del un cubo **copia** con la carpeta local**.**

```
PS C:\Users\Mañana\Downloads\JSON\DESCARGA_S3> aws s3 sync s3://copias3-brayan/ C:\Users\Mañana\Downloads\JSON\DESCARGA_
S3\datos
download: s3://copias3-brayan/datos/manifest.json to datos\datos\manifest.json
download: s3://copias3-brayan/datos/keys.json to datos\datos\keys.json
PS C:\Users\Mañana\Downloads\JSON\DESCARGA_S3> ls


    Directorio: C:\Users\Mañana\Downloads\JSON\DESCARGA_S3


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
d-----        12/01/2026     10:23                datos
```

13. Haz al menos tres consultas SQL cualesquiera sobre algunos de los archivos CSV o JSON que hayas subido.

```
PS C:\WINDOWS\System32> aws s3api select-object-content --bucket copias3-brayan --key datos/keys.json --expression "SELE
CT * FROM S3Object s" --expression-type 'SQL' --input-serialization '{"JSON": {"Type": "DOCUMENT"}}' --output-serializat
ion '{"JSON": {}}' Consulta1.json
```

```
PS C:\WINDOWS\System32> aws s3api select-object-content --bucket copias3-brayan --key datos/keys.json --expression "SELE
CT s.producto FROM S3Object s" --expression-type 'SQL' --input-serialization '{"JSON": {"Type": "LINES"}}' --output-seri
alization '{"JSON": {}}' Consulta2.json
```

```
JSON received: {JSON: {Type: LINES}}
PS C:\WINDOWS\System32> aws s3api select-object-content --bucket copia-tu-nombre --key datos/datos.json --expression "SE
LECT * FROM S3Object s WHERE s.precio > 50" --expression-type 'SQL' --input-serialization '{"JSON": {"Type": "LINES"}}'
--output-serialization '{"JSON": {}}' Consulta3.json
```

14. Elimina los objetos de **primero** de uno en uno.

```
PS C:\WINDOWS\System32> aws s3 rm s3://primeros3-brayan/datos/keys.json
delete: s3://primeros3-brayan/datos/keys.json
PS C:\WINDOWS\System32> aws s3 rm s3://primeros3-brayan/datos/manifest.json
delete: s3://primeros3-brayan/datos/manifest.json
PS C:\WINDOWS\System32> aws s3 ls s3://primeros3-brayan/datos/
```

15. Elimina todos los objetos de **copia** de forma recursiva.

```
PS C:\WINDOWS\System32> aws s3 rm s3://copias3-brayan/ --recursive
delete: s3://copias3-brayan/datos/keys.json
delete: s3://copias3-brayan/datos/manifest.json
```

16. Elimina los cubos **primero** y **copia**..

```
PS C:\WINDOWS\System32> aws s3 rb s3://primeros3-brayan
remove_bucket: primeros3-brayan
```

```
PS C:\WINDOWS\System32> aws s3 rb s3://copias3-brayan
remove_bucket: copias3-brayan
```

17. Fuerza la eliminación del cubo **sincro** sin vaciarlo previamente.

```
PS C:\WINDOWS\System32> aws s3 rb s3://sincro3-brayan --force
delete: s3://sincro3-brayan/datos/manifest.json
delete: s3://sincro3-brayan/datos/keys.json
remove_bucket: sincro3-brayan
PS C:\WINDOWS\System32>
```

CONTENIDO
APARTADO B

 INTRODUCCIÓN

Para poder trasladar la información de HDFS a S3 haremos uso de la herramienta S3DistCp (s3-dist-cp). Analiza su documentación.

De HDFS a S3

1.- Crea un nuevo *bucket* en S3 y utilizando el comando anterior, copia el archivo `ventas.csv` que tenemos en HDFS y obtuviste en la práctica 7.1 a él.

```
PS C:\WINDOWS\System32> aws s3 mb s3://hdfsbrayan
make_bucket: hdfsbrayan
```

```
PS D:\2025 AI IA\BIG DATA\GitHub\BigData2526\PR_07.3_AWS CLI_S3\ventas> aws s3 cp ventas.csv s3://hdfsbrayan
upload: .\ventas.csv to s3://hdfsbrayan/ventas.csv
PS D:\2025 AI IA\BIG DATA\GitHub\BigData2526\PR_07.3_AWS CLI_S3\ventas>
```

2.- (INVESTIGA) ¿Cómo podrás desde HIVE lanzar las mismas consultas que hicimos en los apartados D y E de la práctica 7.1, pero en esta ocasión sobre ficheros en S3?

Desde HIVE se pueden consultar datos en S3 utilizando y creando tablas externas, tendríamos que usar rutas como s3://bucket/ruta/archivo

Podemos seguir estos 5 Puntos clave para realizar as consults correctamente

1. Uso de Tablas Externas: Debes definir la tabla como EXTERNAL para que Hive apunte a los datos en S3 sin borrarlos al eliminar la tabla.
2. Protocolo S3A: La ubicación del archivo debe usar el prefijo s3a:// (ej. LOCATION 's3a://bucket/carpeta/'), que es el estándar de Hadoop para conectar con AWS.
3. Definición de Formato: Es necesario especificar el SerDe (Serializador/Deserializador) adecuado según el archivo original (JSON, CSV o Parquet) mediante la cláusula ROW FORMAT.
4. Configuración de Credenciales: El entorno de Hive debe tener acceso a las claves de AWS (fs.s3a.access.key y fs.s3a.secret.key) o utilizar roles de IAM si estás en un clúster de EMR.
5. Consulta Directa: Una vez mapeada la tabla, ejecutas el SQL estándar (SELECT, WHERE, GROUP BY) exactamente igual que si los datos estuvieran almacenados localmente en HDFS