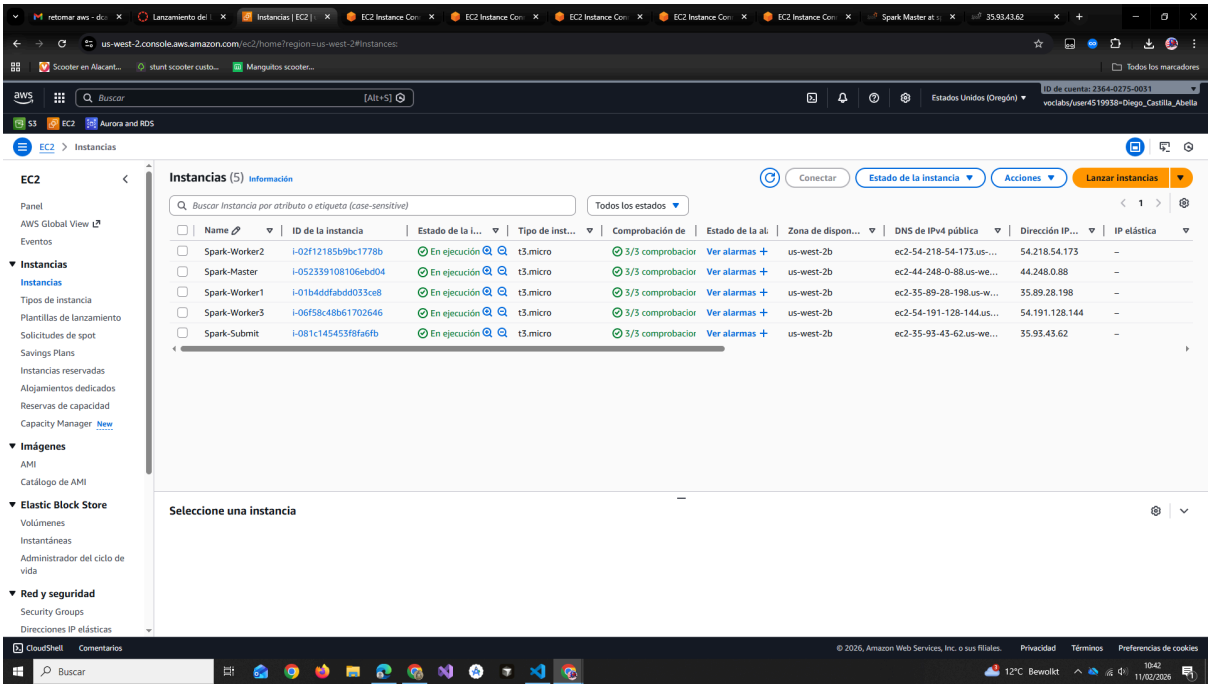
1. Preparación de la Infraestructura en AWS

El primer paso consistió en el despliegue de 5 instancias EC2 independientes para garantizar la separación de roles:

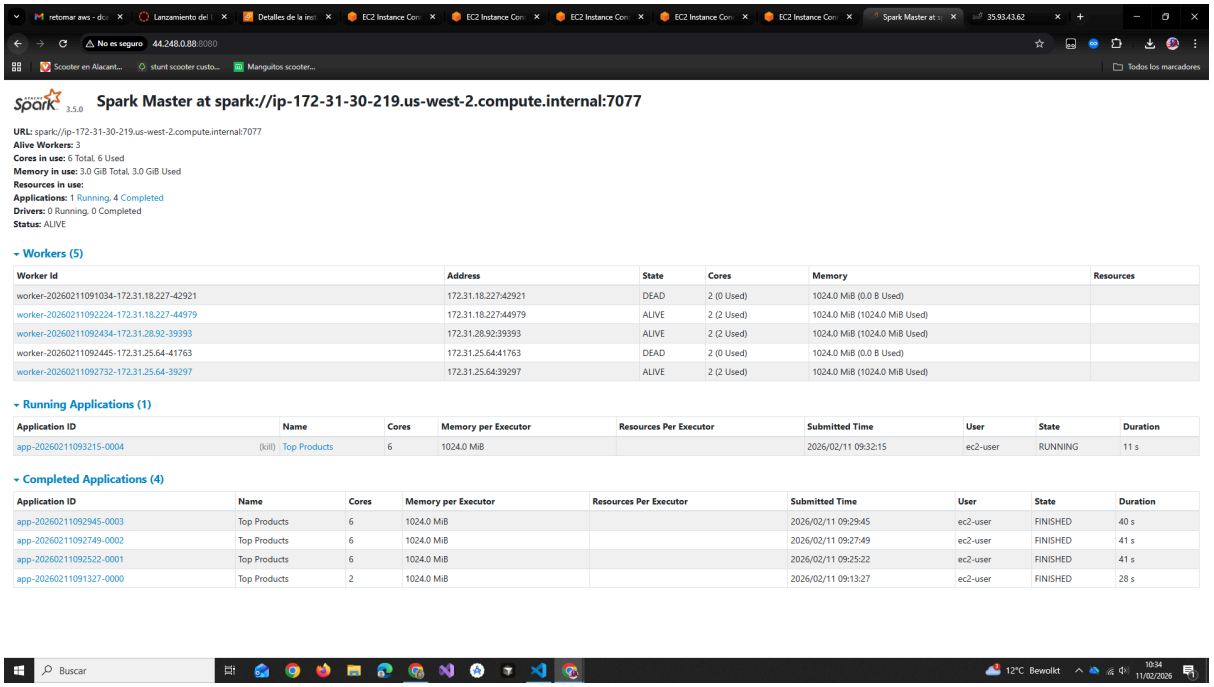
- Spark Master: Actúa como el orquestador de recursos del clúster.
- Spark Workers: Son los nodos encargados de realizar el procesamiento pesado de datos de forma paralela.
- Submit: La EC2 que utilizaremos para lanzar las aplicaciones mediante el comando spark-submit.
- Configuración de Red: Se definieron Security Groups para permitir el tráfico en los puertos 7077 (RPC de Spark), 8080 (Web UI), y 3306 (acceso a la base de datos RDS).



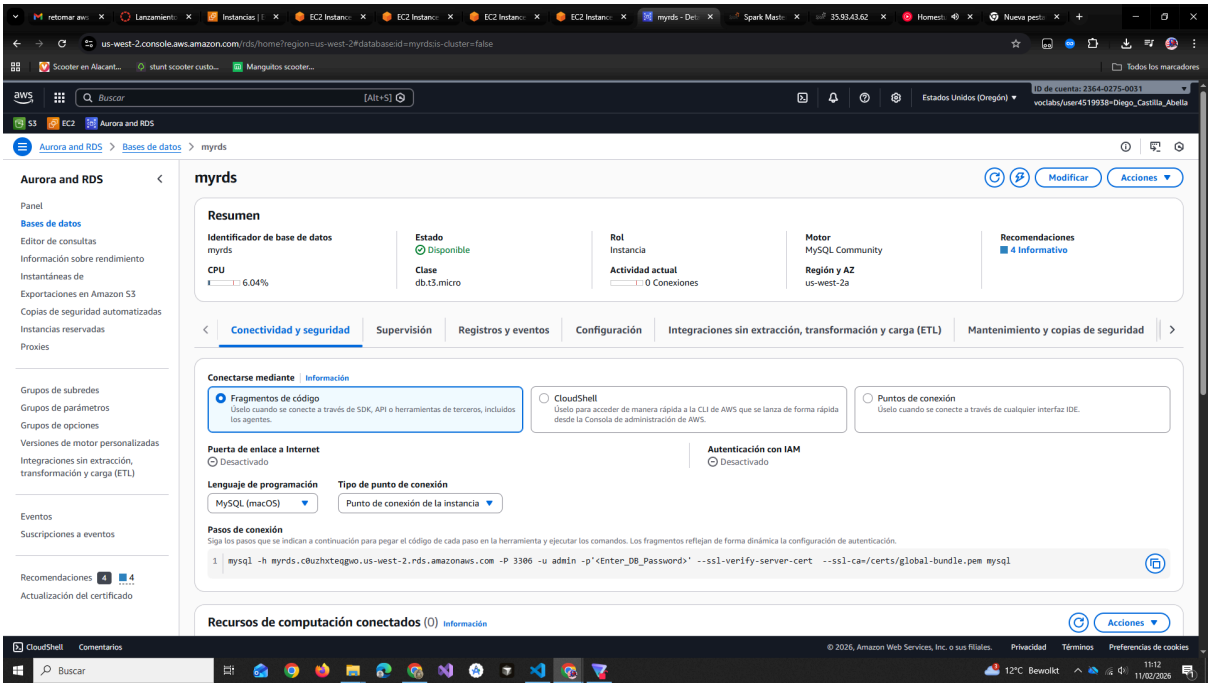
2. Configuración del Entorno y Resolución de Dependencias

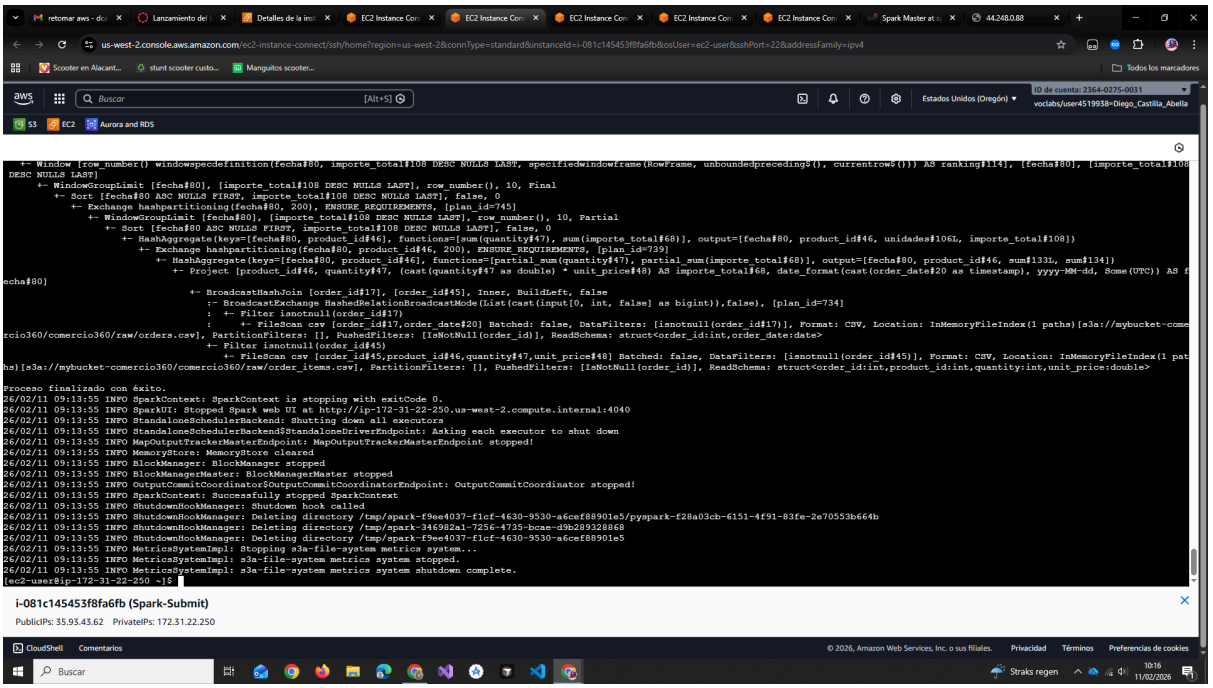
- Una vez operativas las máquinas, se instaló Spark 4.0.1 que al ser una versión tan reciente las librerías de hadoop aun no estaban integradas por lo que tuvimos que hacer el downgrade a la versión 3.5.0 la cual si incluye estos paquetes nativamente, ademas probe primero solo a hacer el downgrade en master y en submit pero entonces los workers no conectaban ya que spark a la hora de comunicarse es muy riguroso con las versiones del mismo. Hacemos downgrade a todo a 3.5.0 y estos ya se conectan y aparecen en el panel web del nodo master 8080.

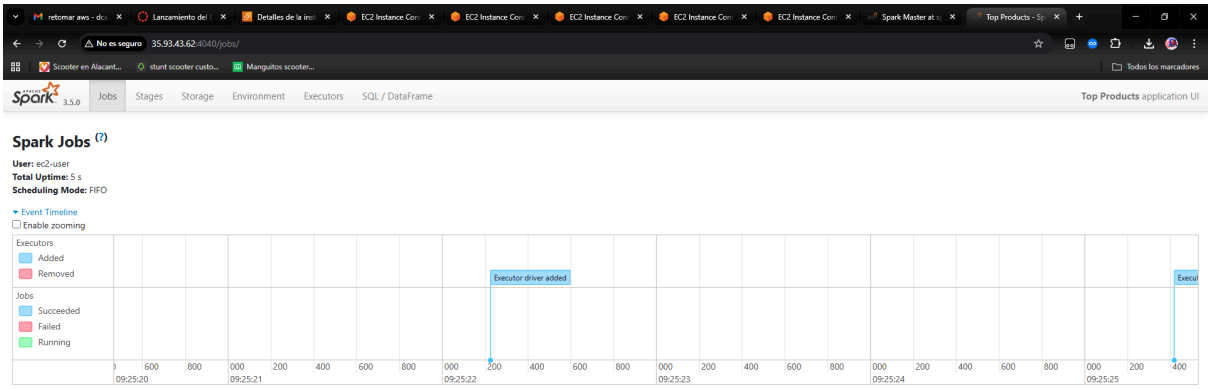
  Una vez con el script python : Spark no podía comunicarse con S3 nativamente. para solucionario se descargaron e integraron manualmente los archivos JAR hadoop-aws-3.3.4.jar y aws-java-sdk-bundle-1.12.262.jar para permitir el uso del protocolo s3a://. Además tambien se incluyó el driver de MySQL para permitir que Spark escribiera los resultados finales en la base de datos RDS.



Diego Castilla Abella


Aurora and RDS
myrds
Modificar
Acciones
Resumen
Identificador de base de datos
myrds
Estado
Disponible
Rol
Instancia
Motor
MySQL Community
Recomendaciones
4 Informativo
CPU
6.04%
Clase
db.t3.micro
Actividad actual
0 Conexiones
Región y AZ
us-west-2a
Conectividad y seguridad
Supervisión
Registros y eventos
Configuración
Integraciones sin extracción, transformación y carga (ETL)
Mantenimiento y copias de seguridad
Conectarse mediante
Fragmentos de código
CloudShell
Puntos de conexión
Puerta de enlace a Internet
Desactivado
Autenticación con IAM
Desactivado
Lenguaje de programación
MySQL (macOS)
Tipo de punto de conexión
Punto de conexión de la instancia
Pasos de conexión
mysql -h myrds.c0uzhxteqgwo.us-west-2.rds.amazonaws.com -P 3306 -u admin -p'<Enter_DB_Password>' --ssl-verify-server-cert --ssl-ca=/certs/global-bundle.pem mysql
Recursos de computación conectados (0)
Acciones

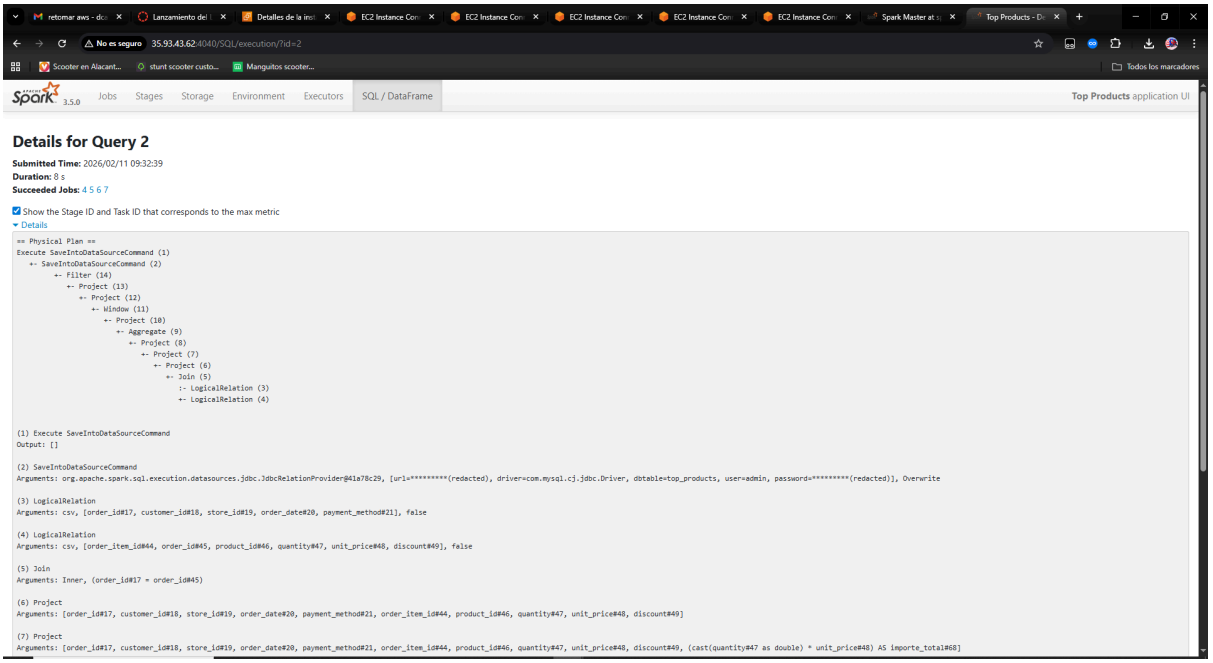

Proceso finalizado con éxito.
26/02/11 09:13:55 INFO SparkContext: SparkContext is stopping with exitCode 0.
26/02/11 09:13:55 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-22-250.us-west-2.compute.internal:4040
26/02/11 09:13:55 INFO StandaloneSchedulerBackend: Shutting down all executors
26/02/11 09:13:55 INFO StandaloneSchedulerBackend$StandaloneDriverEndpoint: Asking each executor to shut down
26/02/11 09:13:55 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
26/02/11 09:13:55 INFO MemoryStore: MemoryStore cleared
26/02/11 09:13:55 INFO BlockManager: BlockManager stopped
26/02/11 09:13:55 INFO BlockManagerMaster: BlockManagerMaster stopped
26/02/11 09:13:55 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
26/02/11 09:13:55 INFO SparkContext: Successfully stopped SparkContext
26/02/11 09:13:55 INFO ShutdownHookManager: Shutdown hook called
26/02/11 09:13:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-f9ee4037-f1cf-4630-9530-a6cef88901e5/pyspark-f28a03cb-6151-4f91-83fe-2e70553b664b
26/02/11 09:13:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-346982a1-7256-4735-bcae-d9b28932886
26/02/11 09:13:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-f9ee4037-f1cf-4630-9530-a6cef88901e5
26/02/11 09:13:55 INFO MetricsSystemImpl: Stopping s3a-file-system metrics system...
26/02/11 09:13:55 INFO MetricsSystemImpl: s3a-file-system metrics system stopped.
26/02/11 09:13:55 INFO MetricsSystemImpl: s3a-file-system metrics system shutdown complete.
[ec2-user@ip-172-31-22-250 ~]$
i-081c145453f8fa6fb (Spark-Submit)
PublicIPs: 35.93.43.62 PrivateIPs: 172.31.22.250



3. Procesamiento de Datos (Flujo ETL)

El proceso de datos siguió el pipeline obligatorio S3 → Spark → RDS:

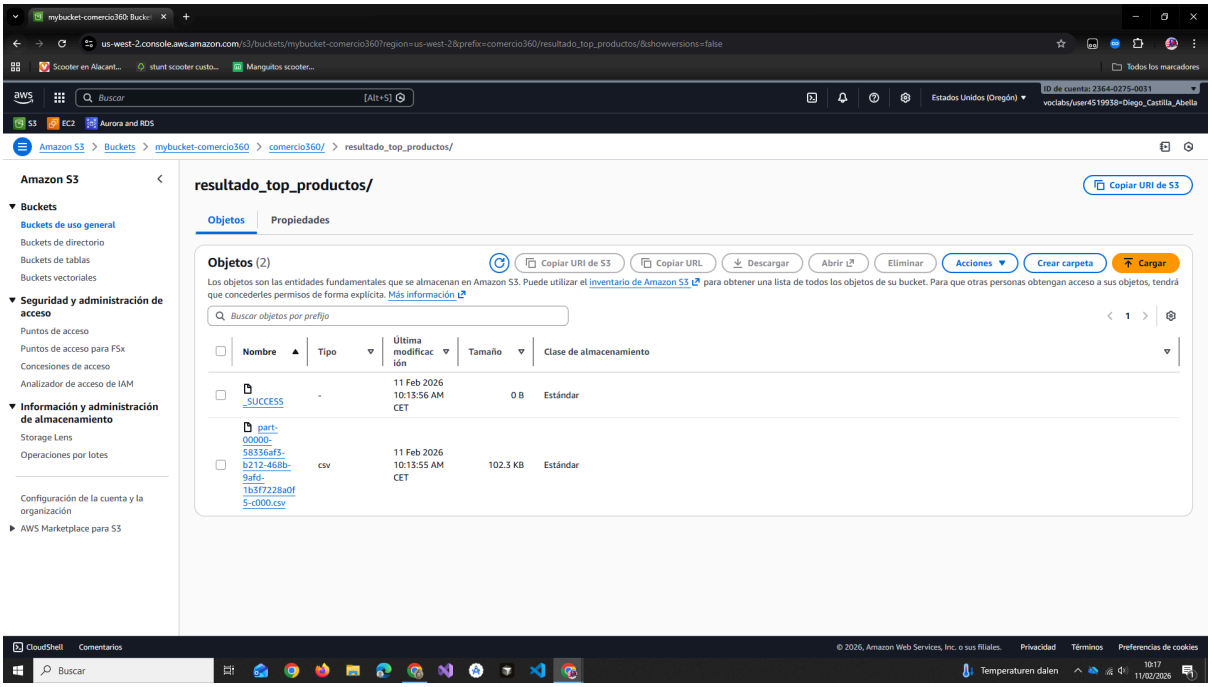
1. **Ingesta** : Spark leyo orders.csv y order_items.csv directamente desde el bucket de Amazon S3.
2. **Procesamiento**:
    - Se realizó un Join entre los pedidos y sus líneas de detalle para asociar productos y precios.
    - Se calcularon las ventas totales por producto y fecha.
    - Se aplicó una función de ventana (Window) para generar un ranking y filtrar únicamente el Top 10 diario por facturación.
3. **Persistencia**: Los resultados se guardaron en S3 en formato CSV para auditoría.
    - Simultáneamente, los datos se insertaron en la tabla top_products de Amazon RDS mediante una conexión JDBC en modo overwrite.

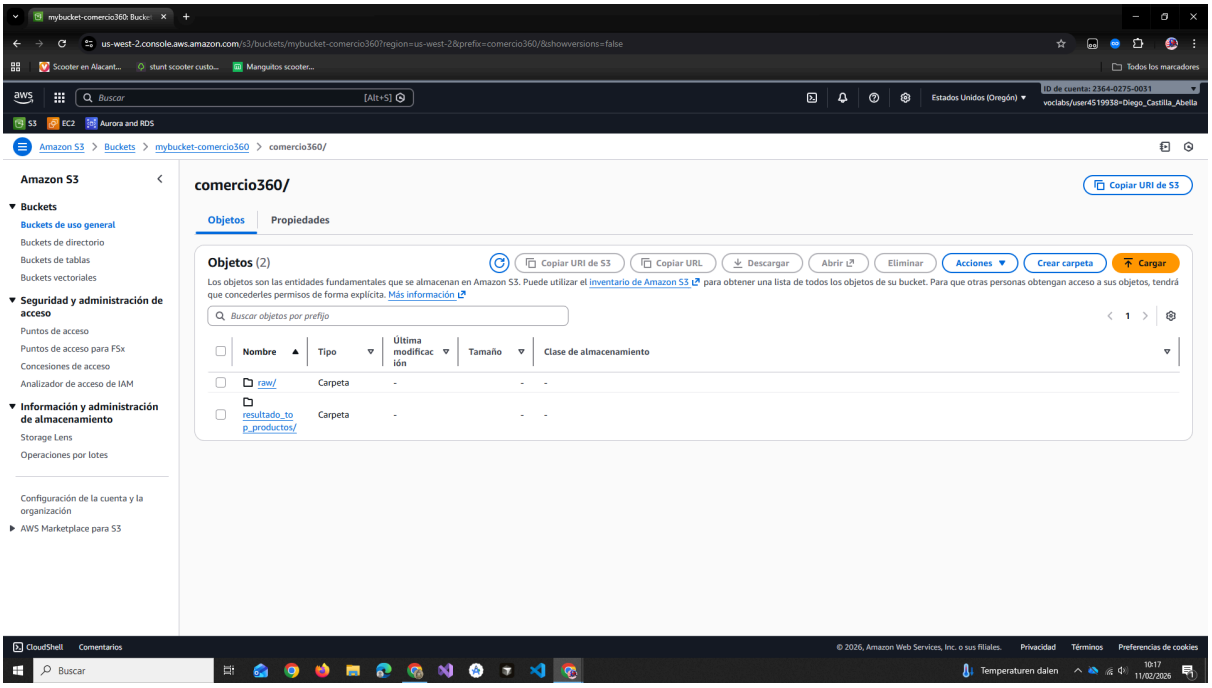


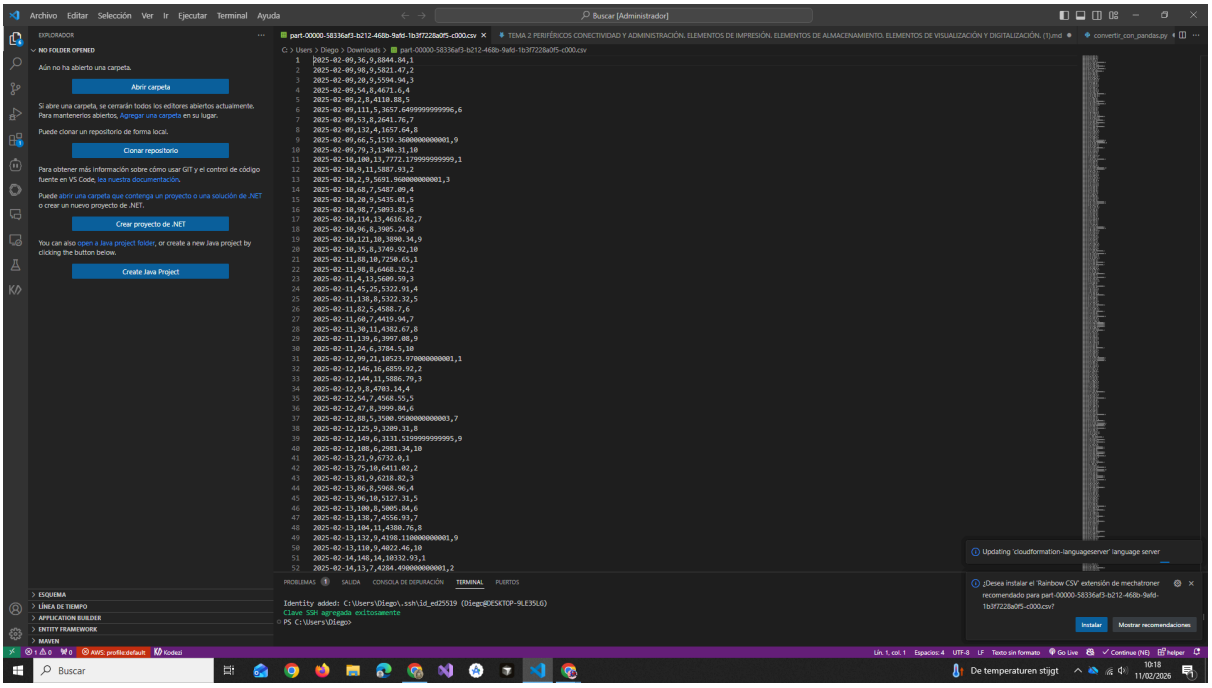
4. Análisis de Ejecución (DAG)

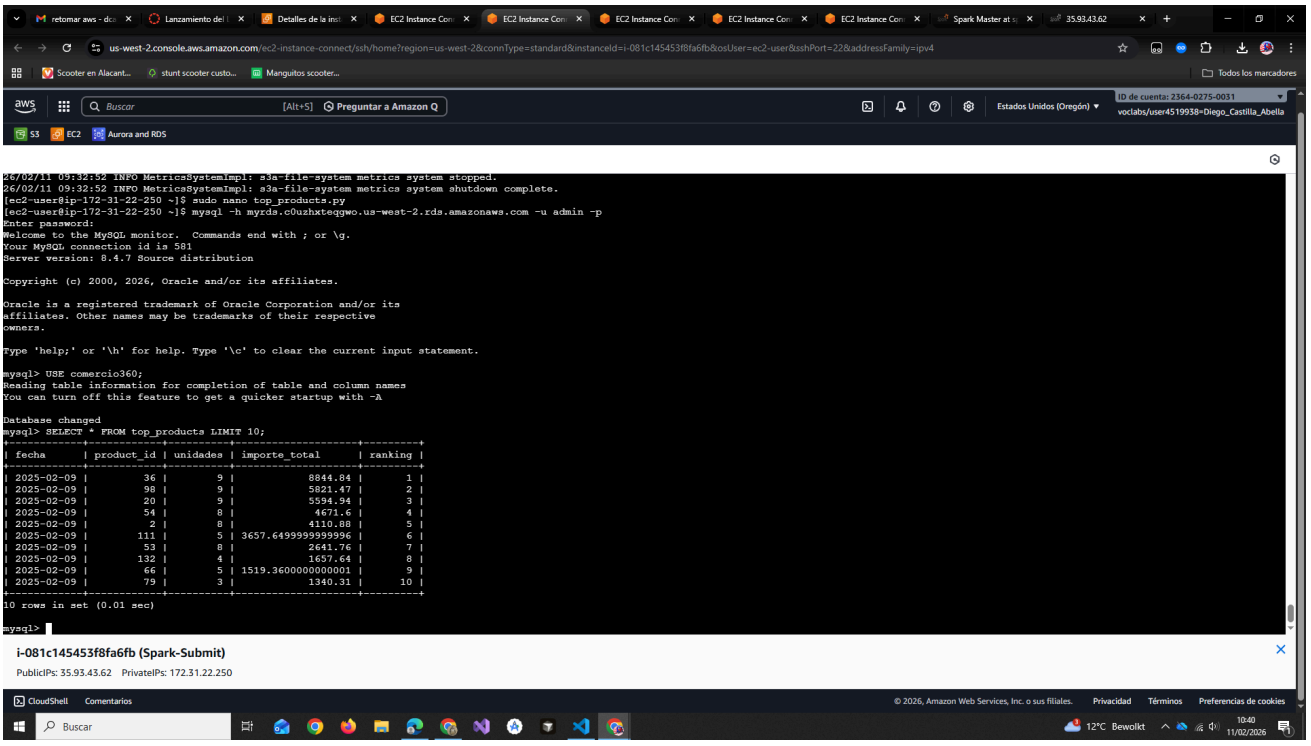
A través de la Spark Web UI, se analizó el Grafo Acíclico Dirigido (DAG) de la consulta:

- El sistema dividió el trabajo en varios Stages (etapas) para optimizar el movimiento de datos entre los workers.
- Se observó un nodo de tipo Exchange para agrupar los datos distribuidos por todo el clúster.
- Finalmente, el comando SaveIntoDataSourceCommand validó que el flujo terminó con éxito enviando los datos a la base de datos externa.


Amazon S3 > Buckets > mybucket-comercio360 > comercio360/ > resultado_top_productos/
resultado_top_productos/
Copiar URI de S3
Objetos
Propiedades
Objetos (2)
Copiar URI de S3
Copiar URL
Descargar
Abrir
Eliminar
Acciones
Crear carpeta
Cargar
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el inventario de Amazon S3 para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. Más información
Buscar objetos por prefijo
Nombre
Tipo
Última modificación
Tamaño
Clase de almacenamiento
_SUCCESS
-
11 Feb 2026 10:13:56 AM CET
0 B
Estándar
part-00000-58336af3-b212-468b-9afd-1b3f7228a0f5-c000.csv
csv
11 Feb 2026 10:13:55 AM CET
102.3 KB
Estándar


Amazon S3 > Buckets > mybucket-comercio360 > comercio360/
comercio360/
Copiar URI de S3
Objetos
Propiedades
Objetos (2)
Copiar URI de S3
Copiar URL
Descargar
Abrir
Eliminar
Acciones
Crear carpeta
Cargar
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el inventario de Amazon S3 para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. Más información
Buscar objetos por prefijo
Nombre
Tipo
Última modificación
Tamaño
Clase de almacenamiento
raw/
Carpeta
resultado_top_productos/
Carpeta





5. Verificación de Resultados

El proceso finalizó en ambos servicios:

- **En S3:** Se generaron los archivos de resultados con la estructura de fecha, producto e importe.
- **En RDS:** Se ejecutó una consulta SQL (SELECT * FROM top_products LIMIT 10;) que confirmó que los 10 mejores productos diarios están correctamente almacenados y disponibles para aplicaciones de negocio.