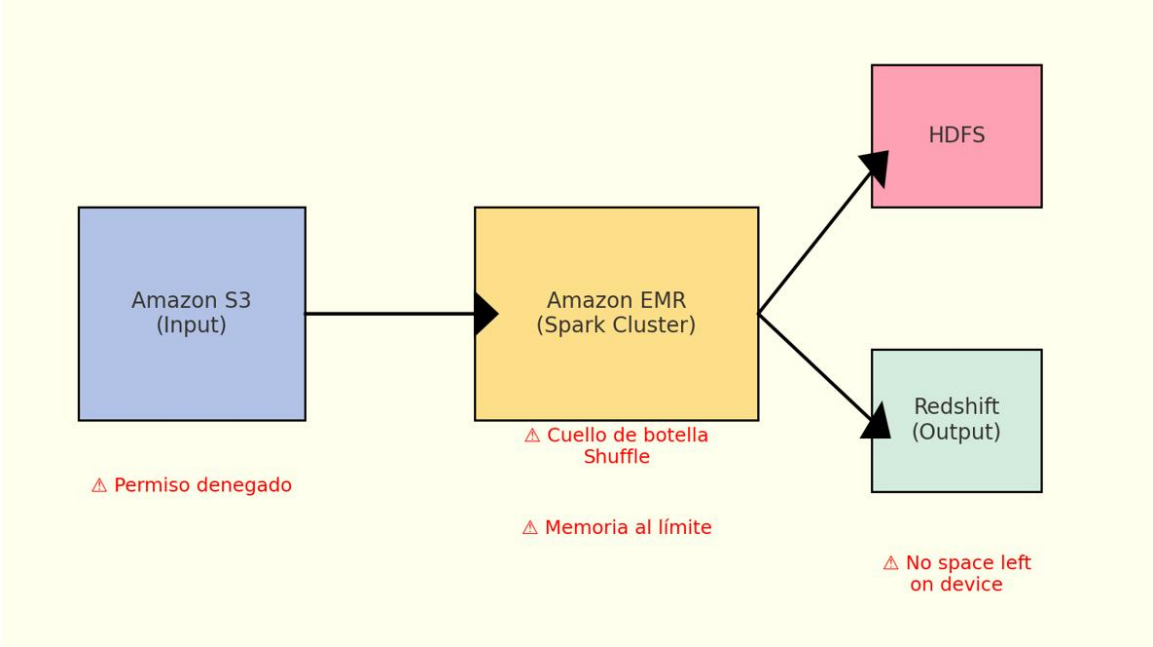
Diagnóstico de errores en un job EMR real

Este documento presenta el diagnóstico de un job de procesamiento distribuido en Amazon EMR. Se analizaron los logs de ejecución de Spark y CloudWatch para identificar errores, sus posibles causas y las acciones correctivas recomendadas.



Problemas detectados y acciones correctivas

Problema detectado	Causa probable	Acción correctiva
Cuello de botella en shuffle (Stage=4, 10GB con 2 particiones)	Muy pocas particiones generan sobrecarga en shuffle	Aumentar spark.sql.shuffle.partitions (ej. de 2 a 50)
Tareas lentas (Stage=2, Task=5, duración 120s)	Mala paralelización o desbalanceo de carga	Optimizar particionado de datos y distribución de tareas
Memoria cercana al límite (Executor=2, 7.8GB/8GB)	Configuración insuficiente de memoria en los nodos executor	Incrementar memoria asignada a cada executor o añadir nodos core
Permiso denegado en lectura de archivo (/user/data/input.csv)	El usuario spark no tiene permisos IAM/HDFS adecuados	Revisar políticas IAM y ACLs de HDFS/S3 para permitir acceso
Fallo en escritura HDFS (No space left on device)	Disco lleno en cluster o mal manejo de particiones temporales	Liberar espacio, mover temporales a S3 o ampliar nodos de almacenamiento

Conclusión

El análisis muestra que los problemas en el job EMR se deben a una combinación de mala configuración del clúster (particiones y memoria), permisos insuficientes y falta de espacio en HDFS. La aplicación de las acciones correctivas permitirá mejorar el rendimiento, asegurar la disponibilidad de recursos y garantizar la correcta ejecución de futuros jobs en EMR.