



# Bases de datos en AWS (RDS, DynamoDB, Redshift)

## </> Índice

- Tipos de bases de datos en AWS
- ¿Qué es AWS Redshift?
- Características de AWS Redshift
- Lab - AWS Redshift

# </> Tipos de Bases de datos en AWS

- **Bases datos relacionales:**

Son las que estamos acostumbrados a usar la mayor parte de nosotros. Existen desde los años 70. Si pensamos en una hoja de cálculo (Excel):

- Base de datos
- Tablas
- Filas
- Campos (Columnas)

itemid	orderid	item	amount
5	1	Chair	200.00
6	1	Table	200.00
7	1	Lamp	123.12

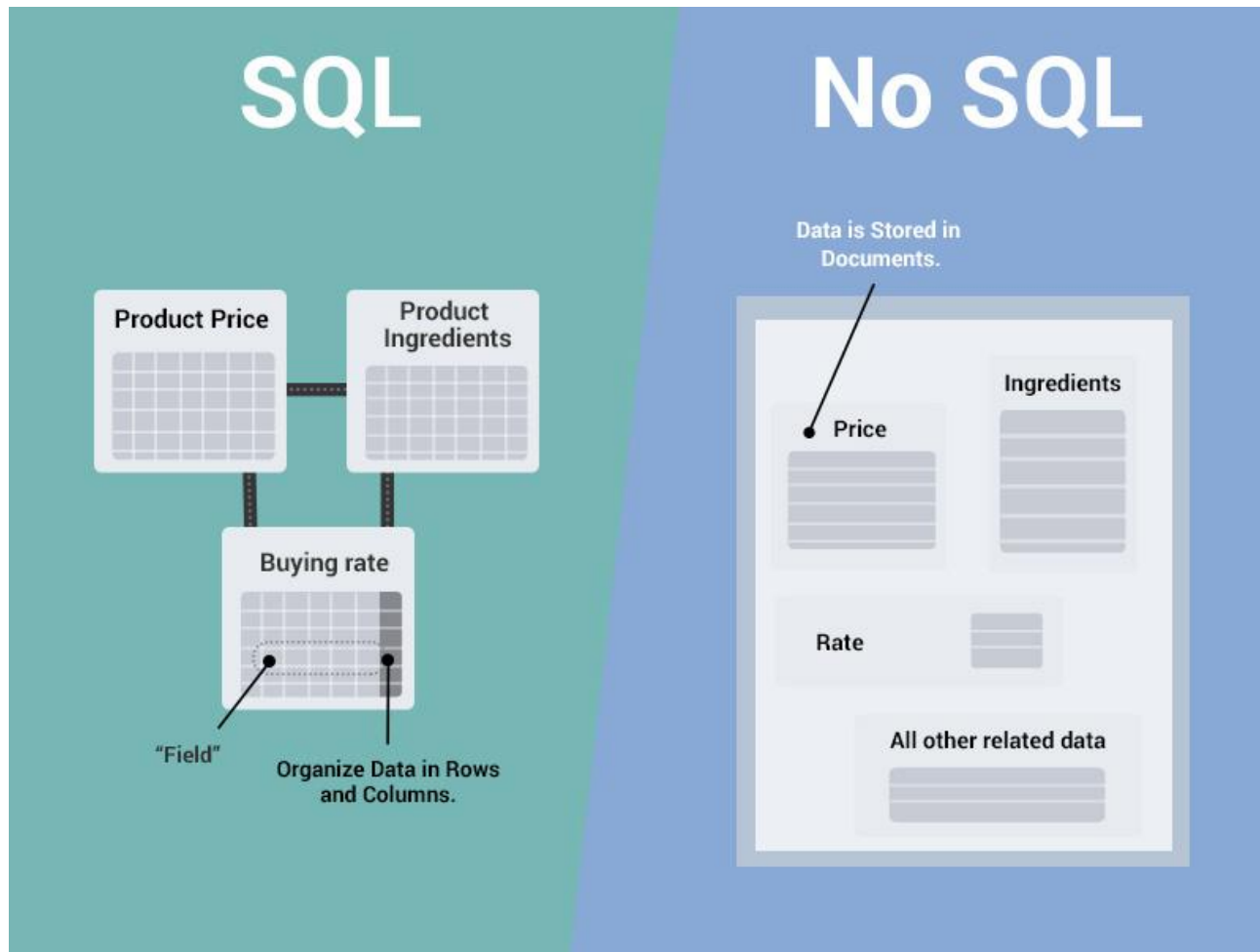
- **Bases de datos no relacionales:**

Las bases de datos no relacionales están diseñadas para modelos de datos específicos y tienen esquemas flexibles para crear aplicaciones modernas:

- Base de datos
- Colección = Tabla
- Documento = Fila
- Pares clave-valor = Campos

Key	Document
1001	<pre>{   "CustomerID": 99,   "OrderItems": [     { "ProductID": 2010,       "Quantity": 2,       "Cost": 520     },     { "ProductID": 4365,       "Quantity": 1,       "Cost": 18     }   ],   "OrderDate": "04/01/2017" }</pre>
1002	<pre>{   "CustomerID": 220,   "OrderItems": [     { "ProductID": 1285,       "Quantity": 1,       "Cost": 120     }   ],   "OrderDate": "05/08/2017" }</pre>

# </> Tipos de Bases de datos en AWS



# </> Tipos de Bases de datos en AWS



Motores de bases de datos RDS en AWS:

- AWS Aurora
- PostgreSQL
- MySQL
- MariaDB
- Oracle
- Microsoft SQL Server



## </> Tipos de Bases de datos en AWS

Motores de bases de datos NoSQL en AWS:

- DynamoDB



Amazon DynamoDB



## ¿Qué es AWS Redshift?

- AWS Redshift es servicio de data warehouse en cloud rápido, potente y altamente escalable.
- Podemos hacer uso de Redshift a pequeña escala desde sólo 0.25\$ la hora o pagar por adelantado si vamos a hacer uso intensivo a razón de 1000\$ por TB por año
- El coste es prácticamente una décima parte comparado con otras soluciones de data warehousing del mercado



## </> Características de AWS Redshift

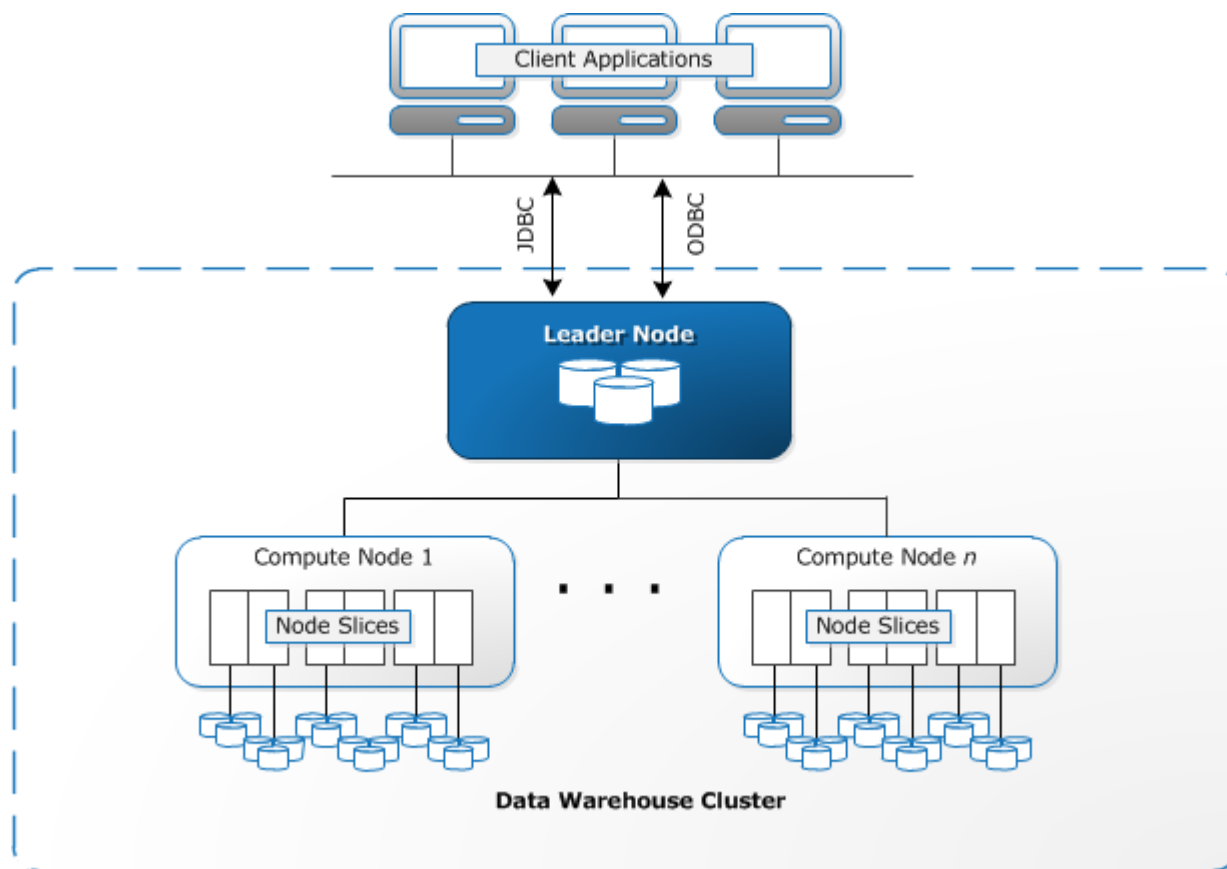
- Procesamiento de datos en paralelo de forma masiva
- Aprendizaje automático
- Almacenamiento en caché de resultados
- Aprovisionamiento automatizado
- Copias de seguridad automatizadas
- Tolerante a fallos
- Almacenamiento seguro y muy escalable



## </> Arquitectura de AWS Redshift

- Un sólo nodo (160GB)
- Multi Nodo
  - Leader Node (gestiona las conexiones de los clientes y recibe las queries)
  - Compute Node (almacena los datos y ejecuta las queries y los cálculos). Podemos desplegar hasta 128 Compute Nodes

# </> Arquitectura de AWS Redshift

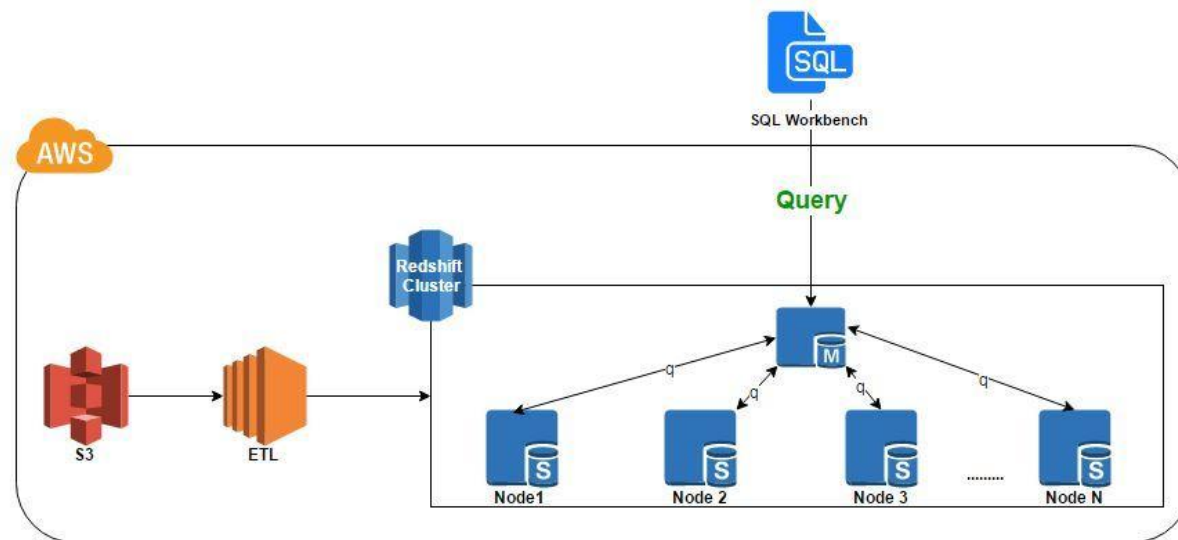


## </> Compresión avanzada en AWS Redshift

- Las bases de datos columnares pueden ser comprimidas mucho más que las basadas en filas debido a que los datos similares se almacenan secuencialmente en los discos
- AWS Redshift emplea varias técnicas de compresión, consiguiendo mejoras respecto a data stores tradicionales
- Cuando cargamos datos en una tabla vacía, AWS Redshift automáticamente muestrea los datos y escoge el sistema de compresión más apropiado

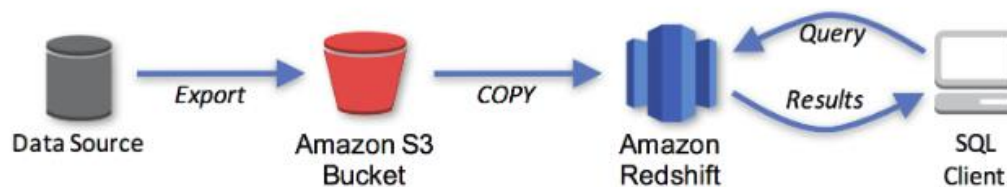
# </> Procesamiento masivo paralelo (MPP) en AWS Redshift

- AWS Redshift distribuye los datos y la carga de las queries automáticamente entre todos los nodos
- AWS Redshift simplifica el escalado y hace sencillo añadir nodos, permitiendo con ello poder mantener el rendimiento de consulta de nuestros datos



## </> Carga de datos en AWS Redshift

- Los datos se exportan desde el sistema de origen de la compañía (una base de datos por ejemplo)
- Los datos se suben a un bucket de AWS S3, si es posible comprimidos para ahorrar espacio
- Los datos que se suben a AWS S3 se copian en tablas de AWS Redshift con el comando COPY
- Utilizamos un cliente SQL para hacer queries a los datos que hemos copiado en AWS Redshift
- Obtenemos el resultado de nuestras queries a través del cliente SQL





## Backups en AWS Redshift

- Están habilitados por defecto con un periodo de retención de un día
- El periodo máximo de retención es de 35 días
- AWS Redshift siempre trata de mantener al menos 3 copias de nuestros datos (el original, la réplica en los compute nodes y un backup en AWS S3)
- AWS Redshift también puede replicar las snapshots de forma asíncrona a AWS S3 en otra región para permitir recuperación de datos ante fallos en una región

## </> Costes en AWS Redshift

- Se cobra por horas de compute nodes
- Se factura una unidad por nodo por hora. Si tenemos 3 nodos en nuestro cluster 24 horas al día 7 días a la semana, se facturarán 2160 horas de compute nodes
- El leader node no tiene coste, sólo se cobra en base a los compute nodes
- Backups
- Traslado de datos (dentro de la VPC)

## </> Disponibilidad en AWS Redshift

- Sólo disponible en una zona de disponibilidad. No podemos desplegarlo en multi AZ
- Podemos restaurar snapshots en una nueva AZ en caso de pérdida de servicio en alguna de las AZs





# ¿Cuál es la mejor manera de aprender sobre bases de datos en AWS?

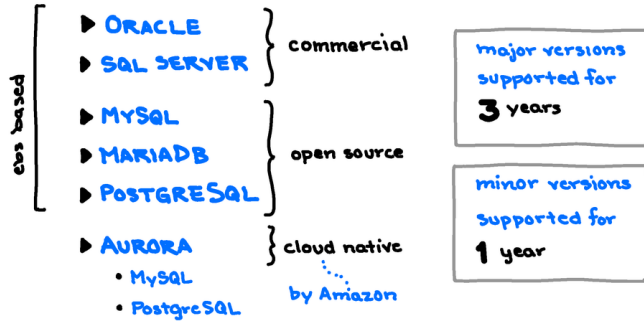
*"For the things we have to learn before we can do them, we learn by doing them"*

*Aristóteles*

### ¿Qué hemos aprendido?

- Cómo crear una base de datos MySQL en AWS RDS
- Cómo configurar el security group creado por defecto para poder acceder a la base de datos desde otras instancias
- Cómo conectarnos a la base de datos que hemos creado
- Cómo borrar la base de datos creada

## ENGINES

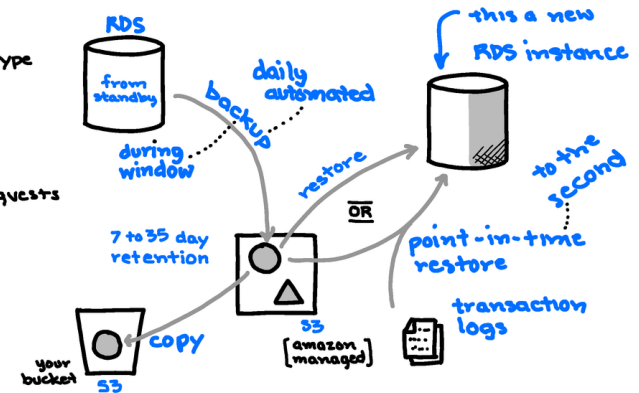


## PRICING

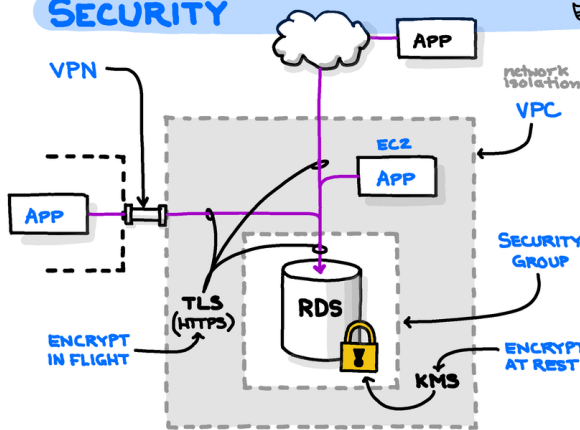
COSTS: (Oregon)

DB INSTANCE HOURS	per engine/instance type billed hourly, round up
GP2 ...	11.54/GB/Mo
STORAGE	101 ... 12.54/GB/Mo + 104/10/Mo
	Mag ... 104/GB/Mo + 104/10/Mo Requests
BACKUP STORAGE	100% of DB size free 9.54/GB/Mo after
DATA TRANSFER	varies by destination

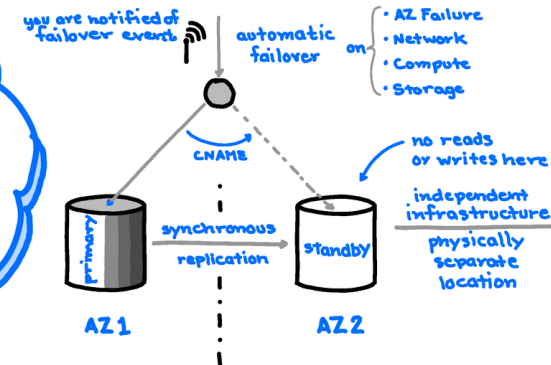
## BACKUPS



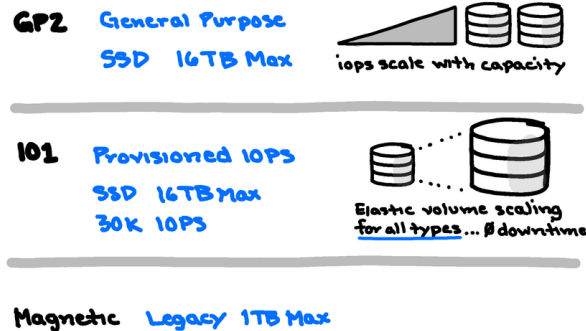
## SECURITY



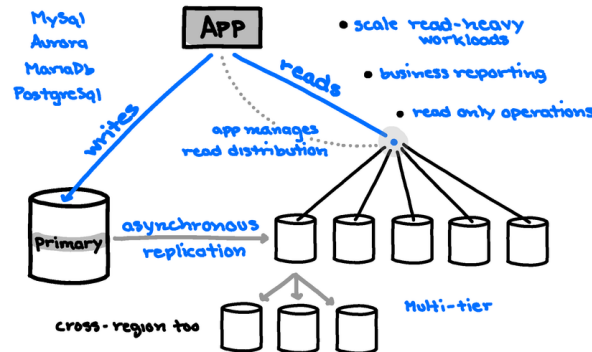
## MULTI-AZ



## STORAGE



## READ REPLICAS



## INSTANCES

### SMALL WORKLOADS

<b>T2</b> burst	1 vCPU	8 vCPU
	1 GB RAM	32 GB RAM

### CPU INTENSIVE WORKLOADS

<b>M3/M4</b> general purpose	2 vCPU	64 vCPU
	8 GB RAM	256 GB RAM

### QUERY INTENSIVE WORKLOADS

<b>R3/R4/X1</b> mem optimized	2 vCPU	128 vCPU
	16 GB RAM	3904 GB RAM

## </> Consejos - RDS

- Si queremos usar MongoDB o Cassandra tenemos que buscar alternativas
- Las instancias RDS tienen por defecto zona UTC

## </> Importante - RDS

- Tienen las restricciones de volúmenes EBS
- Hay que evaluar la BBDD escogida porque igual no tiene todas las funcionalidades que necesitamos
- Si usamos RDS con failover nuestros clientes tienen que atender al failover por DNS que nos aporta RDS
- Limite de 6TB



## Lab - AWS Redshift

### ¿Qué hemos aprendido?

- Cómo crear un clúster de AWS Redshift
- Cómo conectarnos a un clúster de AWS Redshift
- Cómo crear una tabla en AWS Redshift
- Cómo copiar datos de AWS S3 a una tabla de AWS Redshift
- Cómo hacer queries en AWS Redshift
- Consultar métricas en la interfaz de AWS Redshift



## Consejos - Redshift

- Muchas herramientas de BI son compatibles y se pueden integrar con AWS Redshift
- Podemos tunear AWS Redshift para mejorar el rendimiento en función de nuestro caso de uso
- Debemos hacer limpieza de tablas (VACCUUM) para mejorar el rendimiento de nuestras queries



## Importante - Redshift

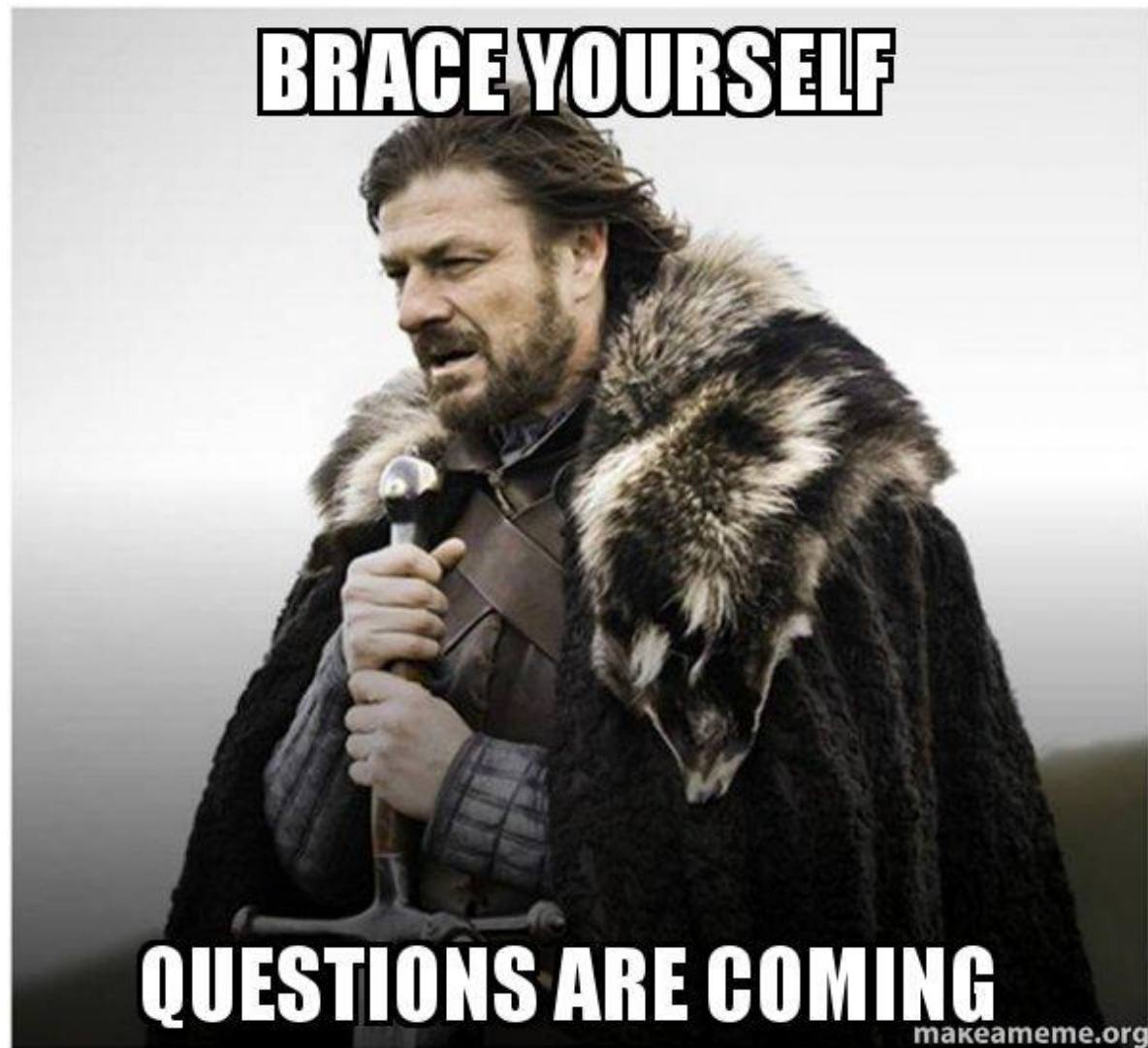
- AWS Redshift no escala horizontalmente. No podemos hacer múltiples queries en paralelo
- El leader node es un punto único de fallo
- El leader node puede hacer de cuello de botella en operaciones que se ejecuten sobre él
- Redshift no soporta despliegues en multi AZ
- Durante la ingesta de datos, el rendimiento de las queries se degrada
- Nunca se debe escalar un clúster que esté en funcionamiento. Lo suyo es crear uno nuevo a partir de snapshots del original, escalarlo y pasar a usar este nuevo clúster



### ¿Qué hemos aprendido en este módulo?

- Conocer la diferencia entre bases de datos relacionales y no relacionales
- Conocer los tipos de datos que nos ofrece AWS
- Conocer qué es AWS Redshift
- Aprender a usar servicios de bases de datos en AWS


## </> Preguntas





- Unai Arrien
- Email de contacto: *unai.arrien@gmail.com*

info@devacademy.es 

687374918 

@DevAcademyES 