

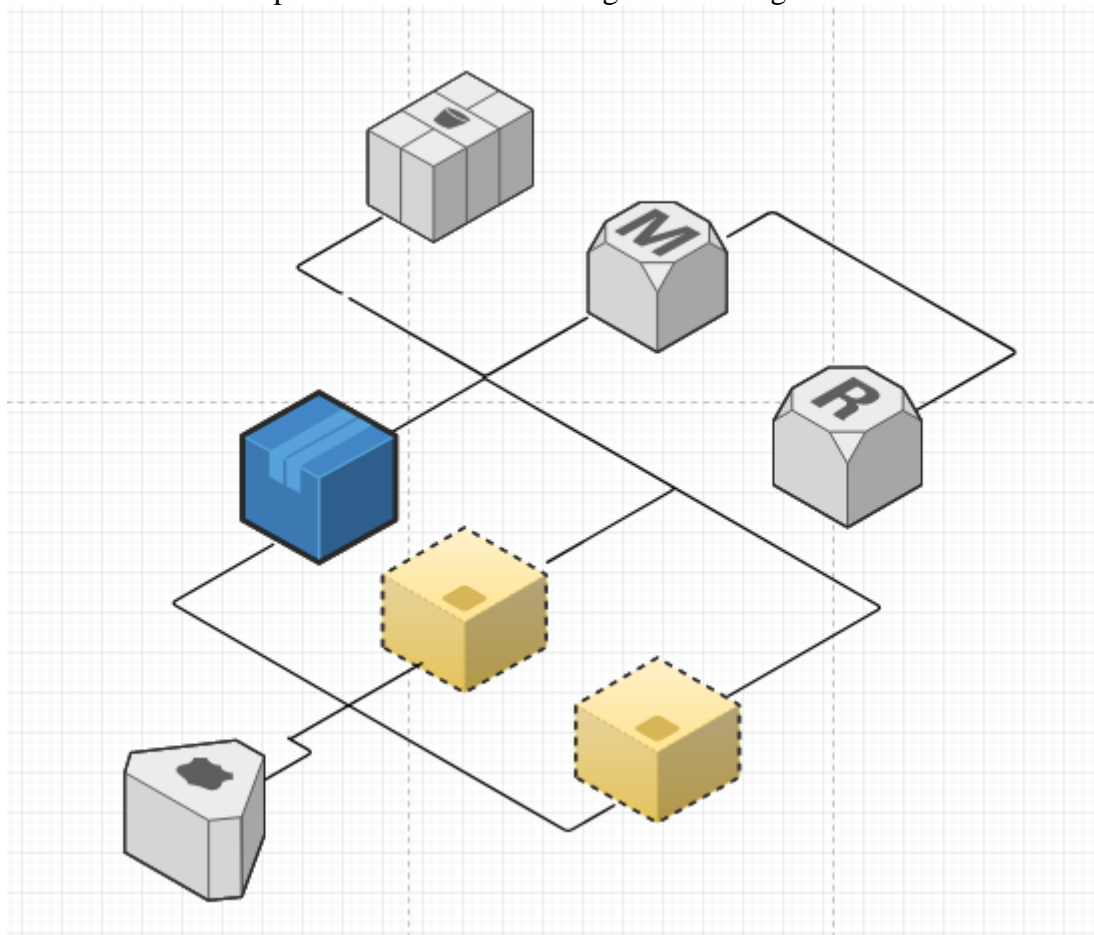
Actividad 1: Distribución de Conjunto de Datos

Se te plantea el siguiente escenario: Trabajas para una empresa de soluciones informáticas para la nube. Tu cliente desea disponer de un servidor web para ofrecer un servicio de descarga de contenido. En concreto, quiere comenzar ofreciendo la posibilidad de que usuarios externos se descarguen un conjunto de datos que permiten recrear un grafo de colaboración entre los personajes de Marvel (sic), disponible en [Marvel Universal Social Graph]. El cliente te comenta su preocupación por tener una solución de bajo coste pero desea poder satisfacer un posible aumento (previsto de antemano) en el número de accesos a la web, sin necesidad de costosas reconfiguraciones. Describe qué solución o soluciones podrías utilizar para implementar la solución que te pide el cliente. No es necesario llegar a realizar el despliegue de prueba.

Como solución al escenario planteado se muestra la siguiente configuración:

Dada

los

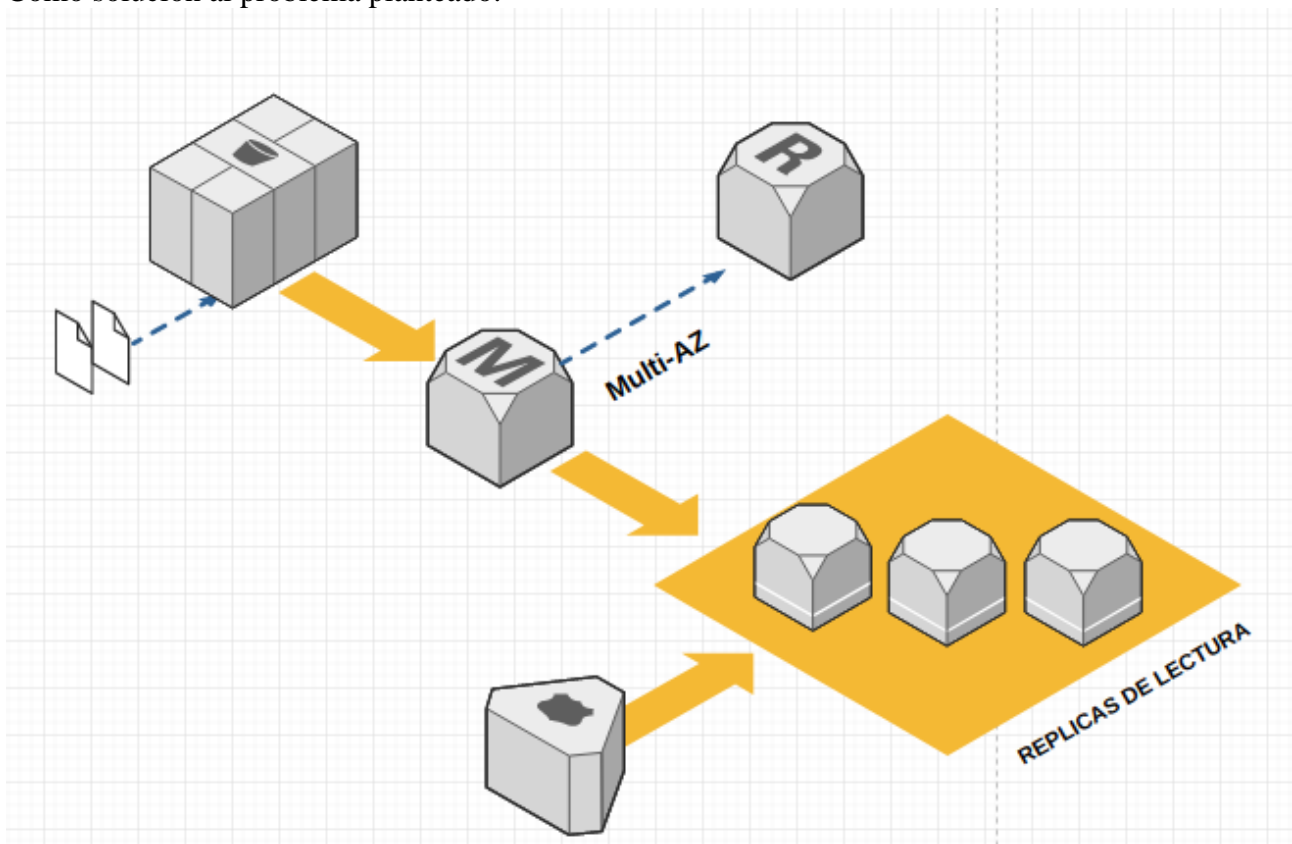


requerimientos se plantea una de las posibles soluciones usando el servicio de almacenamiento S3 de Amazon el cual ira acompañado por un sistema de base de datos el que tendrá su replica. Montaremos en principio una instancia EC2 básica en la cual correrá la aplicación web de tipo Stateless para el registro y monitoreo de usuario el cual se hará uso del servicio Route53 de Amazon para hacer balanceo de carga a nivel de DNS y evitando los costos de un balanceador de carga. A medida que la demanda del sitio incrementa se agregaran nuevas instancias EC2 de aplicación y se registraran en el Service Router53.

Actividad 2: Gestión de Datos en la Nube

Se te plantea el siguiente escenario: Trabajas como Data Scientist de un banco y te entregan una colección de datos como un gigantesco dump de MySQL (un fichero .sql que contiene una secuencia de comandos válida para recrear una base de datos). Te piden que dichos datos deben estar disponibles para ser consultados desde una base de datos y que, además, se cumplan los siguientes requisitos: - La base de datos debe soportar incrementos en las operaciones de lectura (SQL SELECT) en unos determinados días de la semana, sin que esto afecte a las prestaciones de otro tipo de operaciones. - La base de datos debe ser tolerante a fallos sin que haya prácticamente downtime, de manera que el acceso a los datos pueda seguir siendo posible en caso de fallo de una instancia que aloje la base de datos. - No se desea incurrir en gastos de inversión en equipamiento ni en software específico de gestión de base de datos. Razona el esquema que utilizarías y justifica tus decisiones. Puedes incluir un diagrama si lo estimas oportuno (no es necesario realizar un despliegue real).

Como solución al problema planteado:

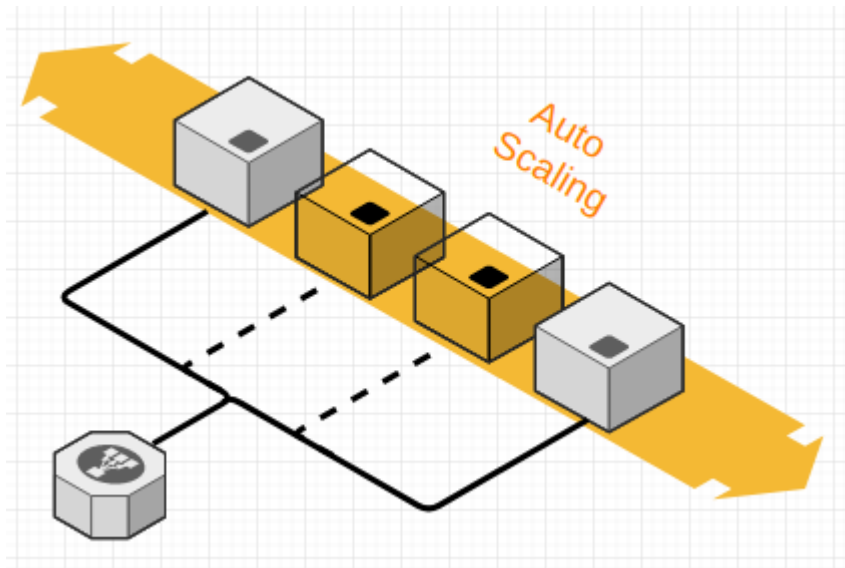


mediante el uso de S3 haremos la carga de los archivos del cual se descargarán en nuestro servidor de base de datos, este servidor contara con el servicio de replicación multi-AZ para evitar las caídas del servicio además agregaremos un sistema de réplicas de lectura del servidor de base de datos al que se le agregara el servicio de Router53.

Actividad 3: Alojamiento de Datos en la Nube

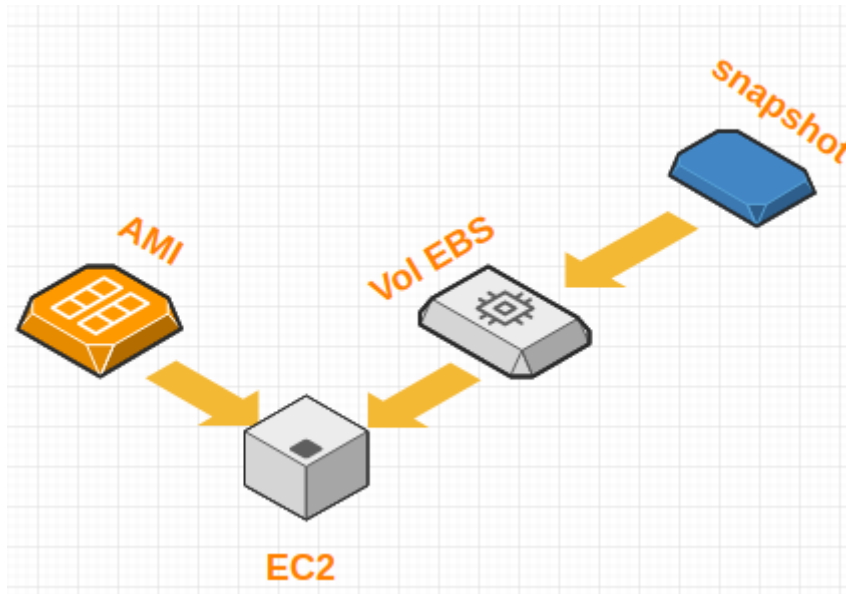
Se te plantea el siguiente escenario: Trabajas en un centro de investigación europeo que necesita gran capacidad de cómputo para analizar una cantidad masiva de datos procedentes de una red de telescopios a nivel mundial. ¿Cuál de las siguientes tecnologías y servicios de Amazon Web Services (AWS) utilizarías para resolver cada uno de los siguientes supuestos? Justifica tu respuesta.

- Supuesto 1. Se necesita capacidad de cómputo extra temporal (mayor número y/o prestaciones de las instancias) para poder ejecutar una aplicación científica durante un cierto espacio de tiempo para obtener los resultados que permitan finalizar a tiempo una publicación científica antes de final de mes.



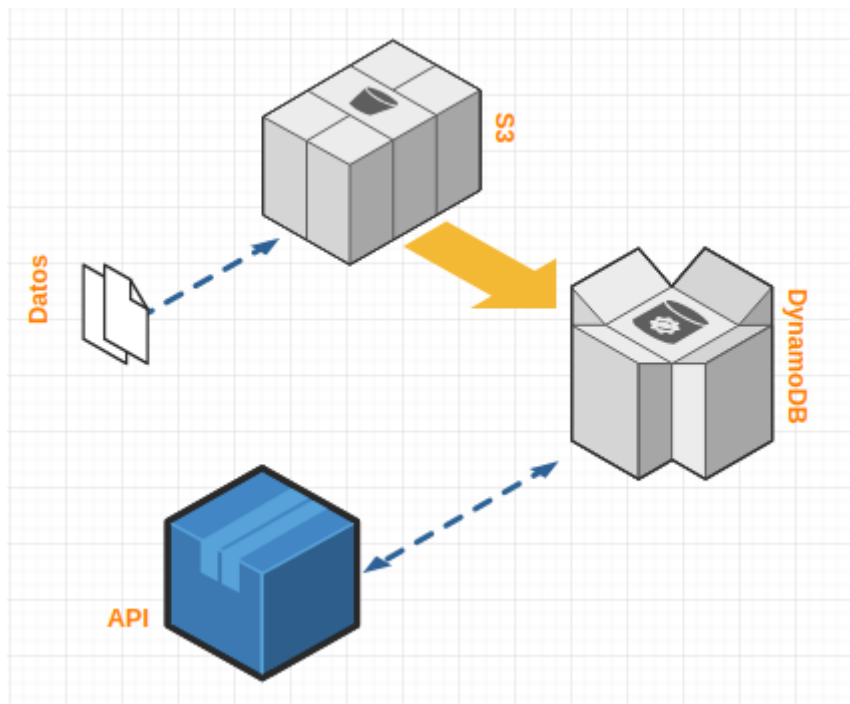
Haciendo uso del servicio de auto scaling de amazon podemos definir políticas de escalado a raíz de las exigencias de recursos.

- Supuesto 2. Se dispone de una AMI (Amazon Machine Image) configurada con una aplicación científica y se dispone de un conjunto de datos público disponible como un snapshot (por ejemplo alguno de estos: https://aws.amazon.com/datasets?_encoding=UTF8&jiveRedirect=1). Se desea que la aplicación en ejecución pueda acceder a dichos datos directamente sin necesidad de introducir ninguna modificación en dicha aplicación.



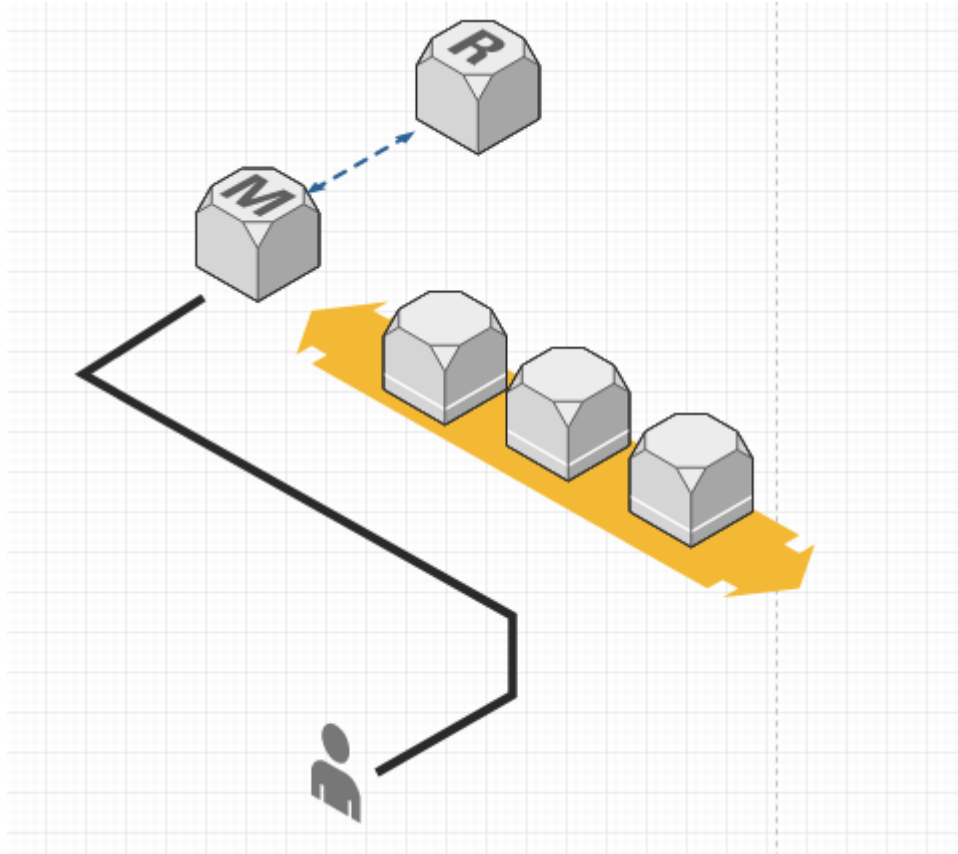
Primero debemos crear a partir del Snapshot de datos un volumen EBS para conectarlo a nuestra instancia de EC2. Lanzaremos nuestra instancia con la AMI configurada y haciendo uso del volumen EBS que contiene los datos de la aplicación.

- Supuesto 3. Se dispone de un conjunto de datos de plantas que no es especialmente grande pero de manera que existen diferentes atributos para cada una de las plantas. Se desea poder consultar dicha información usando un API, teniendo en cuenta que debe ser posible introducir fácilmente información de nuevas plantas que pueden tener nuevos atributos no contemplados previamente.



Primeramente, creamos un servicio S3 para alojar los datos, una vez los datos están en nuestro almacenamiento crearemos a partir de los archivos nuestra base de datos no relacional mediante el servicio de DynamoDB y cuando ya se tengan los datos en base de datos montaremos nuestra API en una instancia EC2.

- Supuesto 4. El centro planea desplegar un nuevo servicio de consulta de un dataset masivo sobre información genómica de múltiples especies. El conjunto de datos está perfectamente estructurado de manera que para cada fila de datos se dispone de exactamente el mismo número de columnas con los mismos atributos. El servicio debe realizar cierto procesamiento de datos relativamente intensivo para decidir la consulta al dataset para obtener un subconjunto de datos. Se desea que el servicio pueda ser escalable porque puede ser muy popular en las siguientes semanas.



Se hará uso del servicio RDS de Amazon con sistema de replicación Multi-AZ y además para soportar alta concurrencia se hará uso del sistema de réplicas de lectura para la consulta de datos masivos.