Propuesta de arquitectura de sistemas de bases de datos no SQL para portal Streaming on line Veeme

Ante la ampliación de de mercado y crecimiento exponencial del número de usuarios del portal Veeme de Elena Streaming video SL se busca una solución para la arquitectura basada en bases de datos No SQL.

Por el presente pliego de condiciones el equipo de Big Data Polska S.L. presenta la siguiente propuesta de solución de arquitectura de bases de datos:

1. Gestión de suscripciones

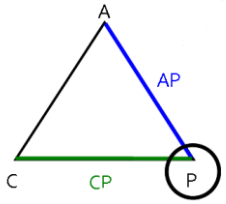
Para el control del acceso de usuarios activos se demanda una alta disponibilidad, siendo menos prioritaria la consistencia de los datos de suscripción de los usuarios.

Los datos de las suscripciones no se actualizan a menudo y no será necesaria una consistencia fuerte, por el contrario si se necesita una gran disponibilidad para que los usuarios puedan acceder a los servicios de Veeme siempre que lo requieran.

Dados estos requerimientos la recomendación técnica es utilizar una base de datos:

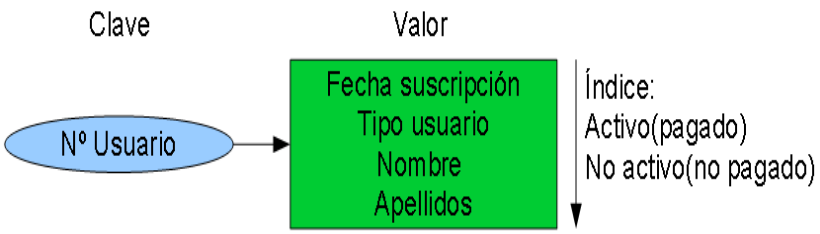
Características ideales de Riak para la gestión de usuarios:

1. - su característica AP prioriza la disponibilidad sobre la consistencia y permite muchos accesos concurrentes.



2. - RIAK permite elaboración de indices adicionales

3. - La información se consultaría en base a clave de usuario



Para asegurar la disponibilidad alta y la consistencia eventual en el tiempo dejaremos la configuración de los valores R y W por defecto (Número de nodos que deben contestar satisfactoriamente para finalizar operación n/2+1).
En cuanto a la gestión de almacenamiento: sloppy quorums garantizan la mayor disponibilidad.

Se descartan las bases de datos como MongoDB por dar prioridad a la consistencia de datos por encima de la disponibilidad, y otras bases de datos de datos clave valor como cassandra dado que no es necesaria una alta escalabilidad de los datos contenidos.

NEo4J da una mayor importancia a la consistencia y al ser una base de datos grafo está más orientada a reflejar las relaciones, algo que no es necesario para la gestión de suscripciones.

2. Catálogo de películas y series

Para la gestión del catálogo de películas y series el requerimiento más importante es la integridad de los datos ya que es la fuente de ingresos principal de Veeme.

Se debe tener como criterio de accesibilidad el país de usuario por cuestiones de copyright. Todos los usuarios deben poder acceder al catálogo por igual, sin embargo solo los usuarios activos podrán ver las películas completas.

En dicha base de datos lo importante son las relaciones:
- información sobre las películas (que están almacenadas en otro servidor de streaming on line) con sus características (año, sinopsis, genero...)
- los actores y su filmografía.
- Otra películas del mismo género

La recomendación técnica para este caso seria:

Las siguientes características de la base de datos Neo4J explican su idoneidad para soportar el catálogo de películas y series:

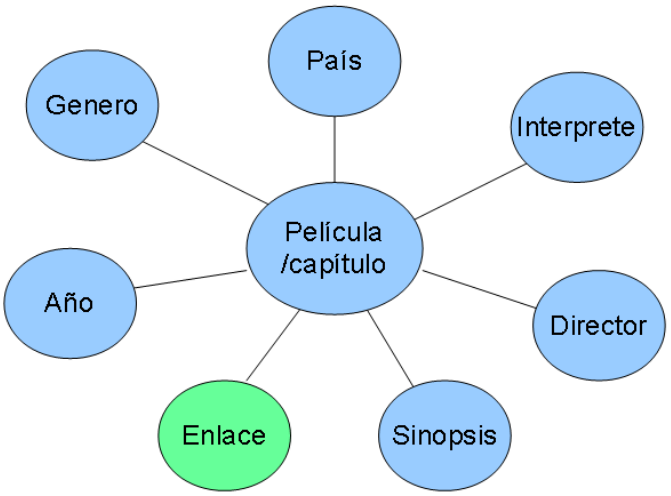
1. Es un modelo de datos en grafo que permite mostrar de una manera optima la relación entre datos como películas, genero, actores…

2. Los tipos de relación entre nodos nos da información adicional (actúa en…, es de tipo…)

3. La tipología de los nodos nos permitirá controlar el acceso de los usuarios a las películas.

4. La escalabilidad es de menor importancia, ya que el número de películas no crece rápido.

Este sería un posible grafo:



Al catálogo se podrá acceder de dos maneras: acceso abierto o introduciendo la clave de usuario. A través de acceso abierto será posible ver el catálogo pero será restringido el acceso a los nodos tipo “Enlace”, a través de los que se establecería el acceso a la plataforma streaming para ver las películas.
El país del usuario se identificaría a través de IP y solo se podrá ver las películas accesibles en el correspondiente país.

Consideramos que un grafo es la mejor de las opciones para representar las relaciones y asegurar la consistencia. Las bases de datos de otro tipo no ofrecen ambas cosas a la vez.

3. Análisis de usuarios

Para extraer información sobre las preferencias de los usuarios hace falta almacenar información sobre cada conexión:

- Hora de la conexión
- Página vista
- Tiempo pasado en la pagina

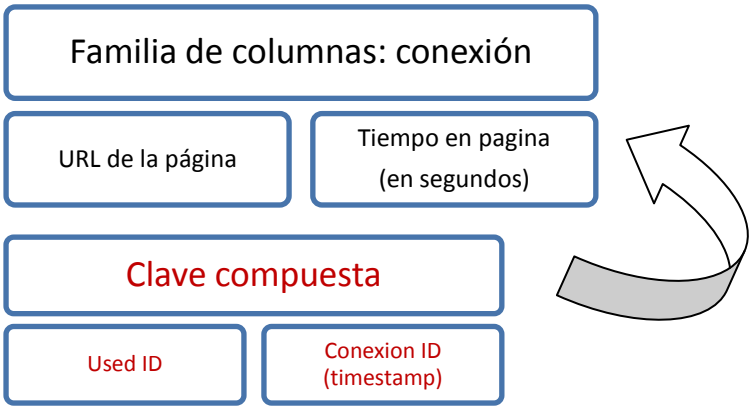
Requisitos: alta disponibilidad y un sistema escalable horizontalmente acorde al crecimiento esperado.

Proponemos almacenar la información en una base de datos basada en columnas, por ejemplo en Casandra, ya que es una base de datos diseñada principalmente para el seguimiento de actividad de usuario y estadísticas.

Siendo un sistema A-P garantiza alta disponibilidad, permite muchos accesos y escrituras simultáneas y soporta gran tráfico de datos.

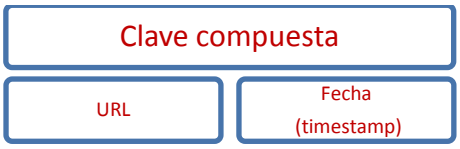
Casandra además ofrece flexibilidad de añadir una información nueva que pueda ser relevante para el análisis de las preferencias del usuario en caso de necesidad futura sin tener que pre diseñar el esquema en el momento inicial.

Los datos podrían estar almacenados en las siguientes columnas:



A partir de las columnas de conexión y su clave compuesta de ID de usuario y ConexiónID se podrá extraer información sobre cada conexión y sobre las conexiones por usuario y URL.

Adicionalmente la base de datos podría contener una columna de contado: Contado con una clave compuesta de: URL y fecha.



El valor de contado se actualizaría cuando ocurriera una conexión. De esta forma se podría extraer de forma rápida las estadísticas de las conexiones a cada URL y agruparlos por intervalos del tiempo.

Para este diseño hemos descartado la base de datos relacionales, documentales y grafos, dada su dificultad en ofrecer disponibilidad alta. Las bases de datos tipo clave valor ofrecen alta disponibilidad al ser tipo A-P, sin embargo nos parecen menos adecuadas por la opacidad del dato y dificultad de consultas complejas.

4. Sistema de recomendaciones ágil

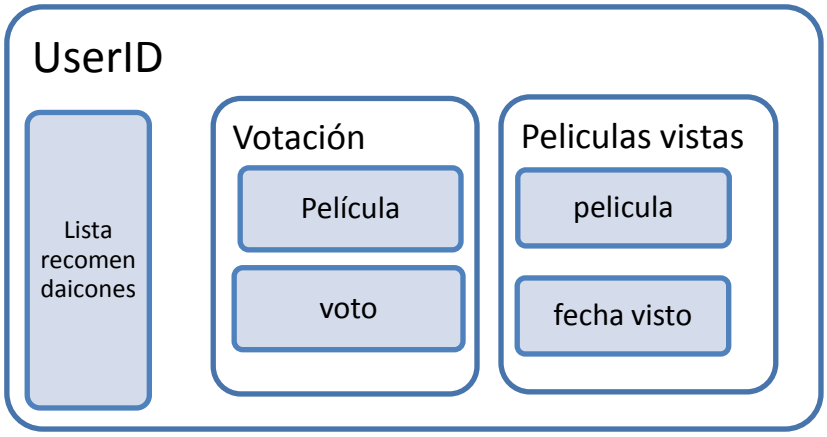
Tenemos el objetivo almacenar la información sobre las recomendaciones en un sistema ágil y accesible de forma rápida. La información a almacenar es de dos tipos:

Las votaciones cada usuario sobre las películas (de 1 a 10) y de las películas vistas últimamente

Recomendaciones ofrecidas a cada usuario de acuerdo a sus preferencias, actualizadas periódicamente

Proponemos como solución una base de datos documental, por ejemplo MongoDB. Su ventaja: buena adaptación a los entornos que requieren escalabilidad es un pro, dado el crecimiento elevado del número de usuarios. MongoDB ofrece muy fácil acceso a los datos a nivel documento que es justo la unidad que se consultaría con más frecuencia.

El agregado de ejemplo tendría la siguiente forma:



Tendríamos un agregado por usuario identificado con UserID. El agregado contendría una lista de las recomendaciones para el usuario correspondiente, un documento incrustado con la lista de películas vistas

Descartamos el uso de las bases de datos basados en grafos, ya que no nos interesa un análisis de las relaciones en este caso, sino análisis del dato en sí. Bases de datos relacionales serían lentas a la hora de hacer consultas en base de joins y no ofrecerían agilidad adecuada. Las bases de datos clave-valor no ofrecerían flexibilidad suficiente a la hora de hacer consultas que no fueran por clave. Aunque en principio las consultas están pensadas en hacerse a nivel usuario, la información a almacenar ofrece más posibilidades de consulta que merece la pena no perder.

Base de datos de columnas ofrecería mejores posibilidades de actualización. Sin embargo esta ventaja se aprovecharía en menor grado, ya que la actualización sería periódica, mientras que el acceso directo y ágil a la lista de recomendaciones por usuario es lo principal.

Equipo Propuesta técnica: 2 personas

- consultora : Anna Lawrenc

- Becario en practicas: Asier Matas