M3. Actividad Colaborativa

Alumno Pablo Olmos Martínez

Índice

1.-Introducción.

En esta actividad colaboradora se nos pide que investiguemos a nivel individual para profundizar nuestro conocimiento de las distintas herramientas Big Data y en concreto de las Herramientas NOSQL; en nuestro caso nos centraremos en la herramienta NEO4J y en su caso de uso para la empresa Adidas.

2.-NEO4J.

NEO4J es un software Opensource de Base de Datos NOSQL orientada a Grafos (BDOG) y programado en JAVA; está orientado para trabajar con grandes Bases de Datos altamente conectados entre sí de una manera eficiente.

Diferentes empresas internacionales utilizan Neo4j con diferentes objetivos:

- Ebay: para la logística de los servicios de entrega (planificación de itinerarios).
- Monsanto: desarrollo y mejores inferencias de los conjuntos de datos genómicos.
- CarterPillar: Neo4j proporciona procesamiento de lenguaje natural a escala, lo que hace que la reparación de equipos sea más eficiente.

Y otros clientes como IBM, Microsoft, Allianz…

3.- Arquitectura NEO4J.

Nos encontramos con 2 tipos de arquitectura:

- **High Available Cluster**: Es la arquitectura que ofrece desde sus inicios Neo4j, que trata de garantizar al máximo la accesibilidad a los datos, aunque varios nodos caigan:

  .-Un clúster está compuesto de una instancia principal (máster) y cero o más instancias secundarias (slave)

  .-Todas las instancias, tanto la principal como las secundarias, tienen una copia completa de toda la base de datos creada con Neo4j.

  .- Todas las instancias se comunican constantemente con el resto para coordinarse y conocer su estado; Las instancias secundarias se comunican con la principal para obtener actualizaciones de los datos.

  .-Las lecturas se pueden hacer sobre cualquier instancia, garantizando la accesibilidad; Las escrituras pueden hacerse directamente sobre la instancia principal o sobre una secundaria.

- **Casual Cluster**: introducida posteriormente, su principal objetivo es dotar de una mayor escalabilidad a la base de datos y una mayor eficiencia de las operaciones de lectura mediante la deslocalización de los nodos, que pueden estar distribuidos en diferentes espacios geográficos. Existen 2 tipos de componentes:

  .- Core Servers: su principal cometido es el de salvaguardar los datos. , siguiendo el protocolo Raft (https://raft.github.io/).

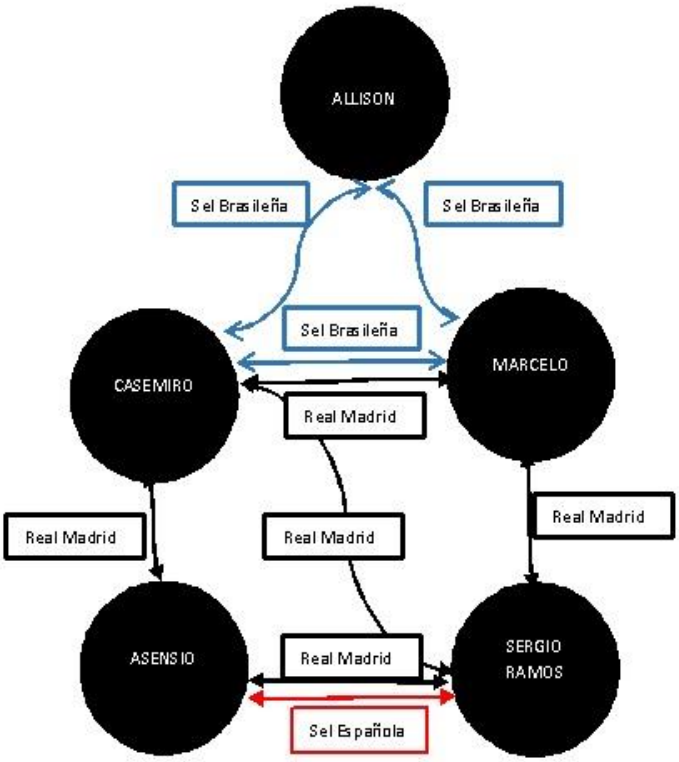
.- Read replicas: el cometido de estas réplicas es el de escalar la carga de trabajo de las operaciones de lectura.

4.- Modelo de Datos.

Siguiendo la Teoría de Grafos NEO4J es una BDOG y se compone de dos elementos:

- **Nodos,** es la unidad fundamental y única sobre la que están constituidos los Grafos, por ejemplo, los futbolistas de un equipo de fútbol.
- **Aristas**, o las relaciones representan las conexiones entre los diferentes Nodos.

Ejemplo, futbolistas del Real Madrid y Liverpool. Los Nodos serían los futbolistas Casemiro, Marcelo, Asensio, S.Ramos y Allison ; y las aristas son el Real Madrid, La selección Brasileña de fútbol y la selección Española de fútbol.



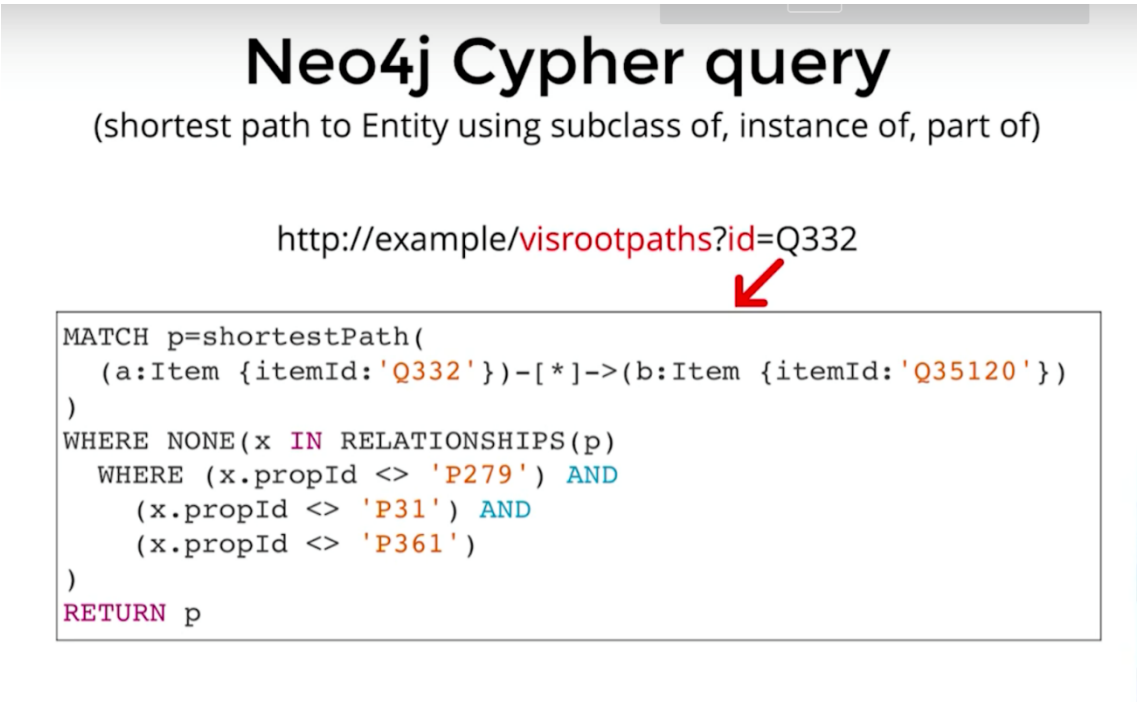
Tanto los nodos como las aristas pueden contener propiedades, que son equivalentes a las columnas de las tablas en el modelo relacional. Y los Nodos pueden tener varias etiquetas que generen nuevas relaciones, Aristas. Por ejemplo, por Año nacimiento, equipos en los que ha jugado, Representante del Jugador, etc…

5.- Criterios de Diseño.

El proceso de modelado consiste en crear una estructura gráfica que exprese las preguntas que queremos hacer de nuestro dominio, por ello se debe:

- Identificar las preguntas del dominio. ¿qué es lo que queremos estudiar?
- Identificar las entidades y las relaciones que aparecen en estas preguntas. Cuáles son nuestros Nodos y Aristas.

Y por último traducir todo esto a **Cypher** es un lenguaje de consulta de gráfico declarativo que permite realizar consultas de datos expresivas y eficientes en un gráfico de propiedades y orientado a la descripción de patrones en grafos.



6.- Ventajas e Inconvenientes de NEO4J.

- Modelo para representar datos conectados. Directamente leíble y fácil de interrogar.
- Implementa consultas referidas a la estructura en grafo gracias al uso de algoritmos basados en grafos.
- Mapeo simple del grafo a lenguajes orientado a objetos como Ruby, Java, C#, Phyton.
- Es altamente disponible (HA) y tolerante a la partición (AP).
- El modelo de datos no está estandarizado dificultad de cambio de gestor.
- Cypher no es un lenguaje estandarizado; Falta de herramientas para ETL.

7.-Business Case NEO4J. Adidas Group Steps up its Game with Neo4j to Personalize the Customer Experience.

Adidas es uno de las marcas más famosas y cotizadas del mundo del deporte, produce cada año más de 600 millones de productos y genera beneficios de miles de Millones de euros anuales.

Para ayudar a impulsar el comercio electrónico, Adidas recurrió a una BDOG Neo4j para explotar las relaciones de datos a través de metadatos compartidos.

Es cierto que el comercio electrónico global abre una amplia variedad de nuevos mercados, pero para su correcta y óptima explotación fue necesario responder de forma rápida (en tiempo real) y adecuada a los distintas necesidades y exigencias de cada cliente.

El objetivo final fue crear un motor de recomendaciones para ofrecer sugerencias relevantes en tiempo real a los compradores en **adidas.com**, además de las **redes sociales** y móviles (**apps**).

El primer y gran problema con el que se enfrentó Adidas fueron las diferentes BBDD con las que contaba:

.-Productos Mercados, Redes Sociales, etc..

, y para llevar a cabo el proyecto de forma exitosa el primer y fundamental paso fue unificarlos y darles el sentido de negocio adecuado al objetivo del Business Case.

En lugar de intentar consolidar todos repositorios de información en un único lugar, Adidas creó **Shared Metadata Service**, logrando el mismo resultado, pero sin la necesidad de mover los datos; esto permitiría a los empleados de Adidas clasificar y buscar contenido en todas las plataformas y divisiones de la empresa, también permitiría dirigirse al público con contenido organizado por idioma, país, tono, deporte y atleta.

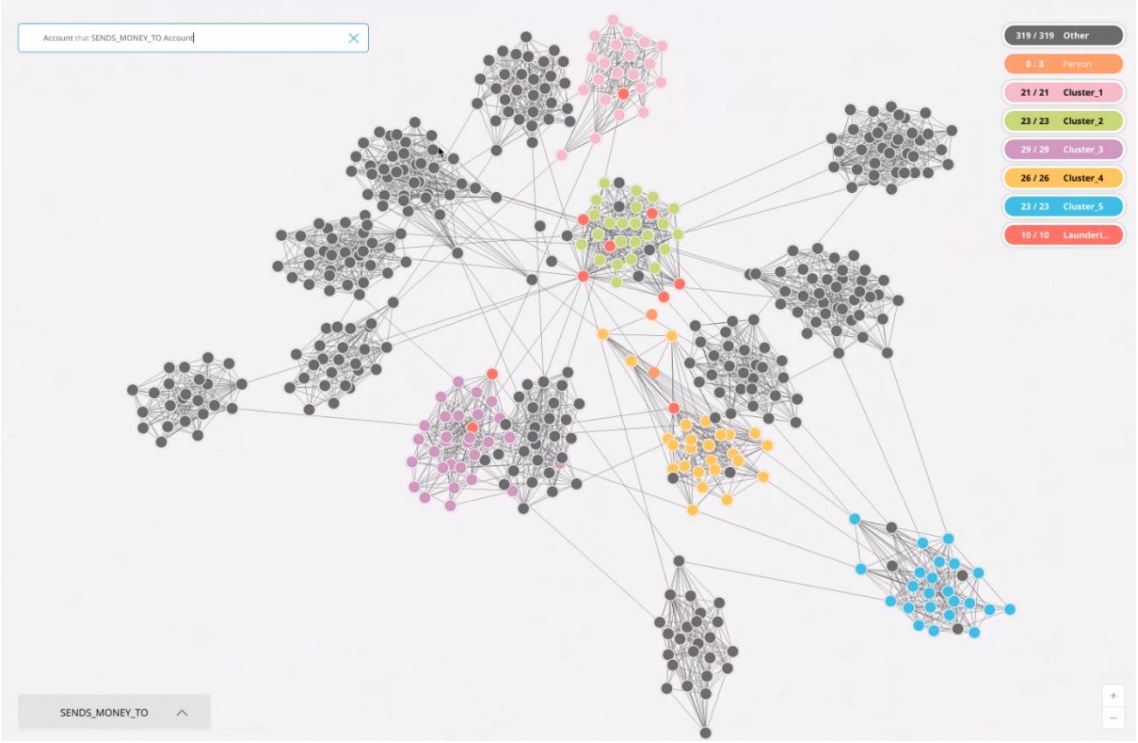
Para construir el **Shared Metadata Service**, Adidas recurrió a **NEO4J**.

La base de datos de Grafos de **NEO4J** permitió el acceso y la búsqueda de todos los datos, junto con el soporte para nuevos servicios emergentes;

El equipo de trabajo de Adidas pudo definir un modelo de datos óptimo que conectaba los tres dominios, relacionando información tan diversa como campañas de marketing, especificaciones de productos, atletas contratados y equipos asociados, categorías deportivas, información de género y más.

Hoy, el **Shared Metadata Service** compartido tiene **2 millones de nodos con casi 10 millones de relaciones (aristas).**

Ejemplo Clusters en Teoría de Grafos con NEO4J



Adidas está integrando continuamente el servicio con nuevas fuentes y clientes, lo que permite al Grupo ofrecer la experiencia online de cliente mejorada adecuada a su segmento.

Por último y gracias al éxito se está trabajando en el motor de recomendaciones en tiempo real, también se están desarrollando perfiles de consumidores basados en jugadores y equipos deportivos favoritos para que el contenido específico se pueda ofrecer al fan apropiado. (Otras integraciones planificadas incluyen el CRM de Adidas, las plataformas de redes sociales y los datos de socios mayoristas y minoristas).

Por ejemplo, un cliente puede navegar y mostrar interés por la NBA, estar en la franja de edad de entre 45-50 años, Geográficamente en Barcelona, con un historial de compra determinado, etc. a nivel grafo estará situado en un Clúster.

Este Clúster a nos da un comportamiento en condiciones Normales dentro de Adidas de cada uno de los clientes y una herramienta para actuar en la relación Cliente-Adidas.

Una representación gráfica que nos ayudará a entender grafo, Vemos que los Nodos son los clientes, Los productos, el Histórico de compra, Los likes/dislikes, Los Iconos (leo Messi) etc…y las aristas la relación de estos nodos entre sí.

8.-Aportación Personal.

Hablar hoy de NoSQL como el futuro de los sistemas de bases de datos puede sonar un poco apresurado, pero hay movimientos importantes con la intención de vencer este miedo al cambio, quizá el mayor problema que todavía pueda existir es la standarización del modelo de datos y la creación de un lenguaje común standarizado de estilo SQL, pero sin serlo.

Como trabajo a futuro había pensado desarrollar gracias a **NEO4J,** el cómo son las relaciones sociales dentro del mundo del fútbol para diferentes objetivos, por ejemplo, como encajaría un nuevo fichaje dentro de mi equipo a nivel Social.

A modo de muy sencillo ejemplo,

Para ello podríamos tener en cuenta:

1.-Toda la BBDD de jugadores de las Ligas Europeas. (NODOS)

2.-Toda la BBDD de representantes de las Grandes Ligas. (NODOS)

3.-Entrenadores (NODOS)

4.-Historíco de equipos en los que ha jugado cada Jugadores (Aristas)

5.-Score para definir el Ratio de calidad de los partidos (Aristas)

Con esta sencilla y más que accesible información podemos construir con **NEO4J** y un algoritmo de Clustering en Grafos diferentes agrupaciones dónde los jugadores se sienten más cómodos, a tenor del Score de Calidad de los partidos.

Estos Clustering nos pueden ayudar a entender si un jugador se va a sentir cómodo en un vestuario determinado porque comparten pasado de otros equipos o comparten experiencia en la selección de su país, o con un tipo de entrenador determinado con un estilo de juego más o menos ofensivo.

Sólo es una información complementaría, pero puede sernos de utilidad en la toma de decisiones.