Universidad Politécnica de Cataluña

Facultad de Informática de Barcelona

Ingeniería de software

Repositorio de árboles genealógicos en BD NoSQL

Author:
Daniel Albarral Nuñez

Supervisor:
Enric Mayol

Q1 - 2015-2016

Contents

1 Estado del arte

Este proyecto se plantea como objeto de estudio y desarrollo la creación de un software orientado a que varios usuarios puedan gestionar arboles genealógicos. El software también ha de ser capaz de trabajar con la datos introducidos por los usuarios, para extraer información que pueda ser útil para ellos. Este tipo de software se llama software genealógico.

1.1 Contextualización

1.1.1 Los arboles genealógicos

Un arboles genealógico, también llamado genorama, es la representación gráfica de los antepasados y descendientes de un individuo.Para su representación se suelen usar tablas o arboles, siendo esta ultima la forma más común y la que se usara en el proyecto.

1.1.2 Uso y aplicación de los arboles genealógicos

Los arboles genealógicos son una herramienta de la genealogía, que se encarga de estudiar y seguir la ascendencia y descendencia de una persona o familia. La genealogía es una ciencia auxiliar de la Historia y es trabajada por un genealogista. Uno de los objetivos del software a desatollar es dar soporte a los genealogistas. Por otro lado hay varias comunidades de aficionados que llevan sus propios arboles genealógicos, el software creado también les podrá dar servicio a esta tipología de usuarios.

1.2 Perspectiva general del software actual.

Todo software genealógico, como mínimo permite almacenar la siguiente información de un individuo: fecha y lugar de nacimiento, fecha de casamiento, muerte y relaciones familiares, contra más flexible es el programa más información te permite introducir acerca de un individuo. También proporcionan diferentes maneras de representar la información y permiten exportar a GEDCOM la información representada.

GEDCOM [1] (**G**enealogical **D**ata **COM**munication): Es un formato de archivo de datos, proporciona un formato flexible y uniforme para el intercambio de datos genealógicos computarizados.

La mayor parte del software genealógico actual esta basado en soluciones de escritorio, pero en los últimos años han aparecido diferentes soluciones web como myheritage o familysearch, que no solo sirven como plataforma de edición sino que también son grandes bases de datos.

Las soluciones más avanzadas en este ámbito aparte de la gestión de arboles también ofrecen herramientas más orientada a la investigación, como podrían ser sistemas de búsqueda de individuos basados en sus relaciones o herramientas estadística.

1.3 Tecnología

Dada la naturaleza social del tipo de software que se busca desarrollar, nacen ciertas complicaciones tecnológicas que en los últimos tiempos se han sido considerablemente investigadas, debido a la gran repercusión de las redes sociales. Las tecnologías clásicas orientadas a la persistencia de datos como las bases de datos SQL, plantean la dificultad de tener un coste muy alto de consulta cuando se pregunta acerca de datos con un alto nivel de relación entre ellos. Una de las soluciones más usadas para solucionar este problema son las bases de datos basadas en grafos, uno de los casos de éxito es el caso de Twitter, que desarrollo su propia solución, FolckDB, una base de datos basada en grafos, tolerante a fallos, diseñada para tratar con grandes conjuntos de datos, con información no critica.

1.3.1 Bases de datos basadas en grafos.

La gran cantidad de proyectos que han nacido en los últimos años hace prácticamente imposible hacer una comparativa concreta de todas las tecnologías existentes. Pero podemos diferenciar el panorama actual haciendo dos generalizaciones:

Tecnologías usadas para propósitos transaccionales .

Orientadas ha dar un servicio online en tiempo real a una aplicación. Estas tecnologías son llamadas bases de datos basadas en grafos. Estas son nuestro principal objeto de estudio. Son el equivalente a las OLTP en el modelo transaccional.

Tecnologías usadas principalmente para el análisis de grafos .

Llamados Motores de procesamiento de grafos, siguiendo el mismo símil que antes, podríamos pensarlos como herramientas de *data mining* y análisis de procesos(OLAP)

OLTP (**O**n**L**ine **T**ransaction **P**rocessing): Es un tipo de procesamiento que facilita y administra aplicaciones transaccionales, usualmente para entrada de datos y recuperación y procesamiento de transacciones (gestor transaccional). Los paquetes de software para OLTP se basan en la arquitectura cliente-servidor ya que suelen ser utilizados por empresas con una red informática distribuida..

Las bases de datos basadas en grafos son sistemas de bases de datos que permiten operaciones CRUD (*Create, Read, Update y Delete*) sobre los objetos representados en ellas. Suelen estar orientadas a funcionar con sistemas transaccionales (OLTP), como aviamos mencionado anteriormente. Sus principales propiedades son, las relaciones son *ciudadanos de primer orden*, a diferencia de otros modelos, como el relacional que tienen que usar claves foráneas. En este tipo de base de datos para representar el dominio de nuestro problema simplemente nos basta con definir los nodos y las relaciones que lo componen.

En el libro Graph Databases [2], encontramos el siguiente gráfico (Figure 1), que nos da una idea de las principales tecnologías basadas en grafos y a que están más orientadas.

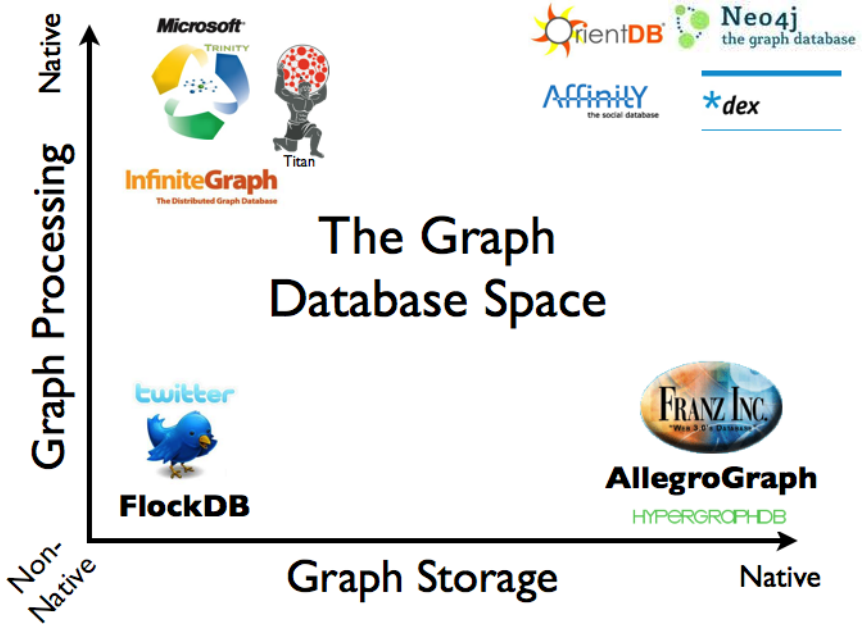

Figure 1: Visión del conjunto de tecnologías

2 Alcance

En este proyecto se desarrolla un software para la gestión y extracción de información de arboles genealógicos. El desarrollo de este software constara de dos fases principales que podemos asilar casi en su totalidad, el desarrollo de una plataforma web para la gestión de los arboles y un sistema que trabaje sobre los datos del sistema, con el objetivo de inferir información relevante del sistema de gestión.

En lo referente a la plataforma web de gestión de arboles genealógicos, constara de un sistema de usuarios que permitirá que estos se conecten a la plataforma para gestionar y almacenar sus propios arboles genealógicos. Los usuarios podrán añadir otros usuarios a sus arboles genealógicos dándoles permiso de edición sobre estos y dando la opción de hacer los arboles públicos, las funcionales concretas se especificaran en las historias de usuario. Este sistema se desarrollara maximizando el aislamiento ente sus componentes para favorecer su mantenibilidad y escalabilidad, con el objetivo de dejar un sistema abierto al desarrollo continuado. Para conseguir este objetivo se trabajara con metodologías ágiles y se dividirá el desarrollo en: Especificación de un modelo conceptual y implantación en una base de

datos basada en grafos (NoSQL) orientada a la optimización de consultas con un alto nivel de interdependencia entre objetos. Una API que ofrezca las operaciones CRUD sobre la BD y gestione el sistema de usuarios y sus *tokens* de acceso. Un *frontend* basado en tecnología web para que los usuarios hagan uso del sistema.

Por otro lado, el sistema encargado de inferir información estará compuesto por demonios que trabajaran usando la API del sistema de gestión. Dado que los objetos que componen la información del sistema de gestión tiene una alta dependencia entre ellos, como ya se ha comentado anteriormente, los demonios tendrán que aprovechar al máximo las herramientas que proporciona la base de datos basada en grafos. La principal función de estos demonios sera encontrar coincidencias entre arboles genealógicos de diferentes usuarios.

2.0.1 Posibles obstáculos

Tiempo: La gestión del tiempo sera uno de los obstáculos más notables a la hora de desarrollar el proyecto, dado que se pretende plantear un proyecto que no necesariamente ha de concluir su desarrollo con la entrega final, para ello se tendrán que plantear claramente las iteraciones necesarias para tener una versión funcional del software y el tiempo de desarrollo que requerirán.

Integración de tecnologías y desarrollo: Uno de los objetivos de este proyecto es combinar ciertas tecnologías con las que no se ha trabajado anteriormente como bases de datos basadas en grafos y *frameworks* JavaScrip para la implementación del *frontend*. Por ello este hecho se tendrá que tener en cuenta en la planificación temporal.

2.0.2 Metodología

El proyecto se desarrollara mediante una adaptación de Scrum que se adecue a las necesidades del proyecto. Los roles se definirán de la siguiente forma:

product owner: El tutor de proyecto asumirá el rol de *product owner* ya que sera el que definirá las historias de usuario y las priorizara.

Equipo de desarrollo y Facilitador: Dada la naturaleza del proyecto el equipo de desarrollo solo sera el alumno que a la vez ha de actuar como facilitador procurando que se cumplan los objetivos de los *sprints*. Por otro lado el *product backlog* se definiría durante la asignatura de GEP.

2.0.3 Herramientas de seguimiento

Dado que el método de trabajo esta basado en Scrum, se usaran las estrategias de seguimiento establecidas por esta metodologías de trabajo ágil. En la fase de desarrollo se harán *sprints* de dos semanas, al principio de cada *sprint* se hará el *Sprint planning* donde se escogerán que historias de usuarios serán desarrolladas. Al final del *Sprint* se hará el *sprint backlog* donde se revisara el trabajo realizado.

References

[1] Ryan Heaton. About gedcom, 2014.

[2] Ian Robinson, Jim Webber, and Emil Eifrem. *Graph Databases*. O'Reilly Media, Inc., 2013.