SOFTWARE HERITAGE

Adrián Alves Morales – UO284288

Luis Manuel Solares García – UO282631

ARQUITECTURA DEL SOFTWARE - Grupo ES2-10

Archivado de software público a gran escala

El archivado de software público a gran escala se refiere a la preservación de software de interés público en grandes cantidades para su acceso y uso a largo plazo.

El objetivo del archivado de software público es hacer que el software esté disponible para la posteridad y garantizar que la historia de la informática sea preservada y esté al alcance de todos, incluyendo investigadores, desarrolladores de software y el público en general.

Acceso y uso de software público archivado

El acceso y uso de software público archivado puede variar dependiendo del proyecto y la institución que lo haya archivado. En algunos casos, el software puede estar disponible para descargar y utilizar de manera gratuita en sitios web de archivo o en repositorios en línea. En otros casos, puede requerirse un proceso de solicitud o incluso una autorización especial para acceder al software. Además, el uso del software puede estar limitado por restricciones de derechos de autor o licencias específicas que deben respetarse.

Software Heritage

Fundada por Roberto Di Cosmo, es una iniciativa para crear un archivo universal de todo el código fuente disponible públicamente con el objetivo de preservarlo, y trata de solucionar los **problemas que surgen** cuando determinados tipos de software son necesarios para:

- Ciencia abierta y docencia, a la hora de realizar investigaciones y restauraciones.
- La industria, que requiere tener un repositorio de referencia de todos los componentes disponibles hoy en día.
- Administraciones públicas, cuando estas intentan mostrar y almacenar software que tiene que ver con la ciudadanía, por ejemplo, para demostrar transparencia. y responsabilidad.

La iniciativa surge a raíz de que muchos de los sitios públicos, tales como GitHub, GitLab, etc., no son realmente archivos ya que, al igual que cualquiera puede añadir un repositorio, al día siguiente ese mismo repositorio puede desaparecer porque su creador lo ha eliminado.

Arquitectura

Debido a la gran diversidad de fuentes de software público que existen, se presenta el problema de la ausencia de estándares a la hora de almacenarlo en dichas fuentes. Cada uno de estos lugares tiene sus propias formas de empaquetar repositorios.

La solución que propone Software Heritage es que la aplicación haga esto automáticamente, de tal manera que, con tan solo hacer uso de una sola línea, de manera similar a la que se haría al hacer cosas como buscar por Google, tendrías acceso a 180 millones de proyectos archivados.

Todos los proyectos almacenados se encuentran en un grafo gigante que los almacena para siempre, por lo que es necesario crear un estándar con un identificador asociado a cada nodo de dicho grafo.

xisten tres metodologías diferentes a la hora de decidir la frecuencia con la que archivar los datos:

- La forma **normal** es mediante un rastreo regular de algunas fuentes, las cuales son distintas unas de otras. Por ejemplo, con GitHub, se realiza un escaneo cada varios meses y se almacenan todos los repositorios que han cambiado en una cola. Esta cola luego es analizada lentamente.
- **Save code now:** Opción para cuando alguien necesita almacenar algún proyecto en el momento, coloca dicho proyecto en la cima de la cola (siempre que provenga de una fuente conocida por la aplicación).
- Realizando **acuerdos** con organizaciones o instituciones para que su software se archive de forma periódica, almacenando también metadatos específicos y control de calidad.

Además, la directiva tiene un acuerdo con GitHub para tener acceso a funcionalidades exclusivas de su API para facilitar el proceso de archivado.

Estructura del grafo

Consiste en un **grafo Merkle**, lo cual permite representar todos estos proyectos y asegurarse de que se puede escalar la aplicación con el enfoque moderno de desarrollo. Este tipo de grafo tiene la capacidad de diferenciar cuando dos contenidos de archivos son iguales, cuando dos directorios son idénticos y cuando dos commits son en realidad iguales. De esta forma, si un archivo o directorio es idéntico o se utiliza en diferentes proyectos, este se mantiene una sola vez. Esto proporciona grandes ventajas a la hora de reducir la cantidad de datos a almacenar.

Desafíos del archivado de software

Posibles desafíos del archivado de software incluyen la selección y priorización de qué software y versiones son importantes de preservar, la elección de la tecnología adecuada para el almacenamiento y acceso a largo plazo, la gestión de la obsolescencia de la tecnología y formatos de archivo, la protección de la privacidad y seguridad de los datos almacenados, y la necesidad de un mantenimiento y actualización constante del sistema de archivado.

Ejemplos de archivados de software público

Algunos ejemplos de archivados de software público vigentes incluyen:

- El Internet Archive, que alberga una amplia variedad de software histórico, como versiones antiguas de sistemas operativos, videojuegos y software de productividad.
- El proyecto GNU, que ha preservado el código fuente de muchos programas de software libre y de código abierto, incluyendo el sistema operativo GNU/Linux.
- La Biblioteca Nacional de Francia, que ha creado una colección de software francés para preservar y difundir la herencia informática francesa.
- El Software Heritage Project, que tiene como objetivo preservar todo el software disponible públicamente en el mundo y hacerlo accesible para la posteridad.
- El Museo de la Informática de Mountain View en California, que cuenta con una colección de hardware y software históricos, incluyendo una amplia variedad de sistemas informáticos antiguos y raros.