Universidad de las Ciencias Informáticas

Facultad 1

"Diseño de arquitectura y configuraciones de seguridad para el motor de búsqueda Orión v2"

Trabajo de Diploma para optar por el Título de

Ingeniero en Ciencias Informáticas

Autora: Betsy Hernández Rojas

Tutores: MSc. Aylín Estrada Velazco

Ing. Paúl Rodríguez Leyva

Ing. Carlos Alberto Robaina

"Año 59 de la Revolución"

"La Habana, junio del 2017"

FRASE

SI QUIERES CONSTRUIR UN BARCO,

NO EMPIECES POR BUSCAR MADERA, CORTAR TABLAS O DISTRIBUIR EL

TRABAJO;

PRIMERO HAZ DE EVOCAR EN LOS HOMBRES EL ANHELO DEL MAR LIBRE

Y ANCHO.

ANTOINE DE SAINT-EXUPERY

DECLARACIÓN DE LA AUTORÍA:

Declaro por este medio que yo Betsy Hernández, con carné de identidad 93092514036 soy la única autora del trabajo titulado Diseño de arquitectura y configuraciones de seguridad para el motor de búsqueda Orión v2¨ y autorizo la Universidad de las Ciencias Informáticas a hacer uso de la misma en su beneficio, así como los derechos patrimoniales con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____

Firma del Autor
Betsy Hernández Rojas

_____	_____	_____
Firma de la Tutora Aylín Estrada Velazco	Firma del Tutor Paúl Rodríguez Leyva	Firma del Tutor Carlos Alberto Robaina

AGRADECIMIENTOS

Este trabajo es el resultado del esfuerzo de muchas personas, que han aportado su granito de arena para que este momento se hiciera realidad. Por tanto, a todas ellas por su apoyo y amabilidad, les dedico este gran logro de mi vida.

Agradecer a la Universidad de las Ciencias Informáticas por ser mi casa durante todos estos años, y ser un pilar importante en mi formación profesional.

Al jurado, por estar presente el día de hoy y dedicarme parte de su tiempo. Muchas gracias

A mis tutores: Paul, Aylin, Carlos, sin su dedicación no hubiera sido posible este trabajo. Especialmente al profesor Paul por estar conmigo desde el comienzo, dedicarme tanta paciencia y comprensión.

A todos mis profesores, en especial a la profesora Yaumara por su entrega y dedicación. Por enseñarnos a no rendirnos ante las dificultades. Muchas gracias

A mis amistades por su apoyo y amistad en cada etapa de la universidad que me ayudaron a ser mejor cada día, a mis compañeros de aula, y en especial a las personitas de la mansión 110-104: Lisbet, Raisa, Wilbia, laura, Baby y Aimet, gracias por su amistad, y por ser mi familia, en tantos momentos de alegría o de tristeza, cuando por motivos de distancia, otros no se encontraban. Además de aguantarme y quererme como soy. Muchas gracias y las voy a tener siempre en mi corazón.

A las amistades y amigas, que, aunque no estén aquí hoy, siempre han estado pendientes de mí y mi carrera. A ellas gracias por formar parte de mi vida. Gracias

A mi novio Carlos, quien ha sido mi amigo, mi compañero, mi compañía, mi fortaleza en momentos
de debilidad, por acompañarme en el paso por esta carrera en los momentos buenos y malos, por su paciencia y comprensión, por ayudarme a ser mejor persona, GRACIAS.

A toda mi familia que siguió de cerca mis estudios en la universidad, a mi hermana por apoyarme y estar siempre cuando la necesite, a mis abuelas y abuelos, por creer en mí. A mis padres, Rossana y Alberto, que con su esfuerzo, cariño, dedicación y fortaleza; me han permitido convertirme en la persona que soy. Gracias

A todas aquellas personas que creyeron en mí, y aún hoy lo siguen haciendo. MUCHAS GRACIAS

DEDICATORIA

Dedico este trabajo de diploma a mis padres Rossana y Alberto por ser los impulsores de mis logros, por sus consejos, por estar siempre apoyándome, por su amor incondicional, por todo lo que han luchado todos estos años porque yo llegara hoy aquí. Por guiarme y enseñarme a ser la persona que soy hoy. Estoy muy orgullosa de ustedes y los adoro.

MUCHAS GRACIAS

RESUMEN:

La presente investigación permitió el desarrollo de un diseño de arquitectura de hardware y configuraciones de seguridad para el buscador Orión v2, sistema desarrollado por la Universidad de las Ciencias Informáticas. Un Sistema de Recuperación de Información (SRI) presenta mecanismos fundamentales como son: rastreo, indexación y visualización. El diseño propuesto cuenta además con nuevos componentes que no se encontraban en la versión anterior, como son: el componente de datos estadísticos, de procesamiento, de apoyo a la toma de decisiones y de datos de usuarios que mejoran el funcionamiento y rendimiento del buscador. La propuesta desarrollada presenta una arquitectura distribuida, permitiendo la escalabilidad de la misma. Para la propuesta se utilizó la metodología AUP-UCI, la arquitectura basada en componentes como estilo arquitectónico, como servidores se presentan Apache Solr (4.10), Nutch (1.9), Nginx (1.10), para el modelado se utilizan Visual Paradim (8.0) y Visio Profesional 2016 y como entorno de prueba la Plataforma de Servicios C.U.B.A. Las pruebas realizadas arrojaron como resultado mejoras en cuanto al rendimiento del SRI. La consulta de expertos permitió validar he interactuar con diferentes puntos de vistas que fueron tomados en cuenta para la propuesta de solución. Como resultado se obtiene la propuesta de diseño de arquitectura de hardware y configuraciones de seguridad para el buscador Orión v2.

Palabras clave: Componentes, Distribuida, Escalabilidad, Hardware, Indexación, Rastreo, Rendimiento.

Contenido

Índice de tablas

Índice de Ilustraciones

INTRODUCCIÓN

En la actualidad la demanda de información produce una profunda transformación social motivada por el desarrollo de las Tecnologías de la Información y la Comunicación (TIC). El mundo está inmerso en un proceso de grandes cambios tecnológicos que permiten estructurar la denominada Sociedad de la Información. Como consecuencia, surgen grandes volúmenes de información, relacionados con documentos, libros, páginas, videos, entre otros, que se encuentran almacenados en la red, ocasionando que pueda resultar engorrosa la búsqueda de la información requerida por el usuario en la web.

La forma más común de acceso a la información alojada en la web es mediante el uso de Sistemas de Recuperación de Información (SRI), programas que permiten localizar coincidencias entre la información que existe en la red y la que demanda un usuario. Estas herramientas funcionan rastreando la red de forma periódica y como resultado de la búsqueda se obtienen aquellos recursos web que se ajustan a los criterios de búsqueda introducidos por un usuario (Pastor, 2016). En este sentido, se destacan los directorios temáticos, los motores de búsqueda o buscadores y los metabuscadores (Pinto, 2015).

Dada la dinámica de Internet en la actualidad los buscadores se caracterizan por la constante indexación de nuevos contenidos en la red. Esto es posible mediante los algoritmos de búsqueda que actúan dentro de un buscador. Las herramientas de recuperación de información o buscadores como son conocidos comúnmente por los internautas, son una fuente de acceso a la información alojada en la web, así como los servicios que en esta se brindan. Su relevancia en el mundo de la navegación por Internet está determinada, dado a que estos son esencialmente rastreadores de información que luego es almacenada y posteriormente accesible a través de consultas, filtros y ordenada según algoritmos matemáticos usados para calcular la relevancia de los resultados obtenidos (Kowalski, 1998).

Los buscadores cuentan con varios componentes que son fundamentales para su funcionamiento. La integración de estos, permite crear una herramienta cuyo objetivo es básicamente rastrear la web, en busca de toda la información que se encuentra dispersa en la misma. Con posterioridad se procede a procesar, almacenar y finalmente brindar interfaces o servicios que permiten a los usuarios consultar la información. El rastreo e indexación periódicos de la web permite que el SRI se convierta en una fuente de almacenamiento constante. Los datos almacenados dependen específicamente del funcionamiento interno del motor de búsqueda y de la estructura real del contenido alojado en el web. El objetivo principal de este componente es rastrear la web y almacenar toda la información encontrada; para lograr esto se basa en dos mecanismos fundamentales: el mecanismo de rastreo y el de indexación

(Maldonado, y otros, 2017). En el caso del componente de rastreo: es el encargado de la búsqueda de información en la web, ya sea documentos, imágenes, videos, entre otras. Estos consultan las páginas web y siguen los enlaces que aparecen en ellas. Pasan de un enlace a otro y recopilan el contenido de dichas páginas que luego proporcionan a los servidores del buscador. El componente de indexación: se encarga del indexado de la documentación rastreada, ya sea para su incremento, actualización o eliminación. El componente de visualización: es el encargado de ser el enlace entre los usuarios y los restantes componentes del buscador, es decir, a través de estos se conectan los usuarios para realizar las consultas y reciben a su vez los resultados. Dichos componentes son la base fundamental de los SRI.

Para un buen funcionamiento de los componentes de un buscador, se analiza la estructura de hardware necesaria y configuraciones de seguridad pertinentes, para lograr el balanceo de la carga y la réplica de los datos. Esto no es más que la distribución de servidores necesarios para rastrear, la indexar y visualizar la información teniendo en cuenta las funcionalidades del SRI que se analice. En el caso de la seguridad de los nodos dentro de la infraestructura, se centra en la aplicación de reglas como mecanismo de control de acceso a través de los cortafuegos. Además, como la aplicación web va a permitir la gestión de usuarios se recomienda utilizar protocolo HTTPS para el uso de la herramienta.

Entre los motores de búsqueda más destacados de internet se encuentran: Google, Baidu, Bing y Yahoo (Suárez, 2016). Estos permiten realizar búsquedas de imágenes, documentos, catálogos, videos, música, noticias y libros (Pinto, 2015). Los mismos cuentan con complicadas arquitecturas de hardware, debido a su gran demanda en la web, además de la variedad de servicios que brindan a los internautas. Estos buscadores en muchos casos limitan algunos de sus servicios para Cuba y no brindan información confiable acerca de la relevancia de los resultados mostrados. Esto hace necesario que sean más los esfuerzos para contribuir en la búsqueda y recuperación de información en los contenidos publicados en la red cubana, con el fin de lograr una mayor disponibilidad de la información científica, cultural e informativa en el país.

En el ámbito internacional hay más de 1 000 000 000 sitios web con contenidos de documentación, noticias, videos, páginas webs, entre otras (Internet live Stats, 2017). Según datos estadísticos publicados por el sitio web del Centro Cubano de Información de Red (CUBANIC), Cuba presenta un total de 6711 dominios .cu en la red, hasta abril de 2017, estos se encuentran distribuidos entre las áreas de cultura, periodismo, salud, política, deportes y otras ofrecidas por sitios webs de distintas índoles. Además, en el área de educación, existen 68 instituciones de nivel superior que incluyen 3150 sedes universitarias municipales que promueven contenidos de gran importancia como son: publicaciones científicas, tesis de pregrado, tesis de maestrías y tesis doctorales, entre otros documentos semejantes

(Cubanic, 2017). Como parte del desarrollo científico alcanzado en estas instituciones se ha generado un gran volumen de conocimiento, en muchos casos documentado en artículos científicos publicados en la intranet cubana.

De este modo se puede apreciar el desarrollo que poco a poco, pese a las limitantes del país en cuanto a las redes de información, se ha podido alcanzar y el gran cúmulo de información que se encuentra almacenada en la red cubana.

En la Universidad de Ciencias Informáticas (UCI), específicamente en el Centro de Ideoinformática (CIDI), perteneciente a la Facultad 1 se desarrolla el buscador cubano llamado Orión v2; cuenta con facilidades para realizar búsquedas sobre páginas web, imágenes y documentos. Contiene componentes como: mecanismo de rastreo, mecanismo de Indexación, mecanismo de visualización y otros que se encuentran en proyección de desarrollo.

Este SRI carece de un diseño de arquitectura de hardware y configuraciones de seguridad necesarias para un software de esa magnitud, pues la estructura que presenta fue implementada para la primera versión de Orión que contaba con un número reducido de componentes. Esto conlleva a que su rendimiento no sea óptimo, ocasionando problemas como: bajo nivel de actualización de la información rastreada, lo cual ocasiona que la información almacenada no siempre sea la más actual, influyendo de forma negativa en el grado de aceptación de los usuarios del sistema según encuestas realizadas a 26 estudiantes que hacen uso periódico de esta herramienta.

Por otra parte, el grado de ralentización de las respuestas brindadas a los internautas, provocado por la ausencia de un correcto entorno de despliegue según lo expresara el jefe del departamento de Servicios Informáticos para Internet (SENIT), donde fue desarrollada la herramienta, ha ocasionado pérdidas de usuarios y a su vez la disminución de los niveles de retroalimentación. Esta situación ha causado que se torne difícil el estudio de criterios positivos o negativos sobre el sistema, aspecto tan importante para la evolución del mismo. Al no tener una arquitectura correctamente estructurada, se complejiza el proceso de asignación de presupuesto para el despliegue del sistema ya que se desconoce qué hardware destinar con estos propósitos. Esta desventaja puede ocasionar gastos innecesarios en la instalación del buscador atentando contra desarrollo económico del centro y la universidad.

Por lo antes planteando se define como **problema científico:** ¿Cómo mejorar el rendimiento y la seguridad de los componentes de hardware del buscador Orión v2?

Como **objeto de estudio** de la investigación: Proceso de análisis y diseño de arquitecturas y configuraciones de seguridad para SRI.

El objeto de estudio planteado anteriormente se enfoca en el **campo de acción:** Diseño de arquitectura y configuraciones de seguridad para el buscador Orión v2.

Para dar solución al problema planteado se propone el siguiente **objetivo general:** Diseñar la arquitectura y las configuraciones de seguridad para el buscador Orión v2, haciendo uso de técnicas que garanticen la seguridad y el correcto rendimiento de este sistema.

Por consiguiente, los **objetivos específicos** son los siguientes:

1. Construir los referentes teóricos fundamentales que sustentan la investigación relacionados con el desarrollo de arquitecturas y configuraciones de seguridad para SRI.
2. Diagnosticar el estado de los SRI en la actualidad en cuanto al diseño de arquitecturas y configuraciones de seguridad.
3. Identificar los requerimientos para el diseño de la arquitectura y configuraciones de seguridad del buscador Orión v2.
4. Diseñar la arquitectura y configuraciones de seguridad del buscador Orión v2.
5. Validar el diseño de la arquitectura y configuraciones de seguridad del buscador Orión v2.

Para dar cumplimiento al objetivo planteado se proponen las siguientes **tareas de la investigación:**

- Realización del análisis del estado del arte de los modelos de Recuperación de Información de los motores de búsqueda.
- Caracterización de las herramientas, tecnologías y metodologías necesarias para el desarrollo de la arquitectura distribuida de hardware para el motor de búsqueda Orión v2.
- Elaboración de la propuesta de arquitectura y configuración de seguridad para el motor de búsqueda Orión v2.
- Diseño de arquitectura y configuraciones de seguridad para el motor de búsqueda Orión v2.
- Realización de pruebas para la verificación del funcionamiento del diseño de arquitectura y configuración de seguridad del motor de búsqueda Orión v2.

Como parte del desarrollo de la investigación, y teniendo en cuenta los objetivos y las necesidades de la misma se formula la siguiente **hipótesis científica:**

Una correcta arquitectura, con las configuraciones de seguridad pertinentes para el buscador Orión v2, mejorará el rendimiento y la seguridad de este SRI.

Se define como **variable independiente:** Arquitectura y configuraciones de seguridad para el buscador Orión v2:

Como **variables dependientes** se presentan:

Variable dependiente 1: Rendimiento

Los indicadores para medir la variable rendimiento son los siguientes:

- carga y estrés

Variable dependiente 2: Seguridad

Para operacionalizar la variable dependiente seguridad, se hace uso de indicadores que se encarguen de medir o comprobar determinadas tareas. Estos indicadores son:

- integridad de los datos: se mide por los errores encontrados, basándose en la protección de los datos, ante modificaciones de personal no autorizado.
- denegación de servicios: se mide los errores encontrados, ante la respuesta de ataques de ante el acceso de la información a los usuarios.
- transición insegura de HTTP a HTTPS: se mide por los errores encontrados, en la información de algunas páginas las seguras.

Operacionalización de las Variables de la Hipótesis Científica

Para el desarrollo de la operacionalización de variables pertenecientes a la demostración de la hipótesis se definen algunos aspectos por variables que son:

Descripción: Describe en que consiste la variable.
Dimensión: Describe el alcance de la variable en la investigación.
Indicadores: Pautas a tener en cuenta por variable para la investigación.

Unidad de Medida: Asigna en un rango numérico.

Tabla 1. Operacionalización de las Variables

Variables	Descripción	Dimensión	Indicadores	Unidad de medida
Independiente	Propuesta de diseño de Arquitectura de hardware y configuraciones de seguridad básicas	Diseño de arquitectura y configuraciones de seguridad para el buscador Orión v2	Diseño de arquitectura	(5) Totalmente de acuerdo (4) De acuerdo (3) Ni de acuerdo, ni en desacuerdo (2) en desacuerdo (1)Totalmente desacuerdo
			Configuraciones de seguridad	(5)Totalmente de acuerdo (4) De acuerdo (3) Ni de acuerdo, ni en desacuerdo (2) en desacuerdo (1)Totalmente desacuerdo
Dependiente 1	Mide los niveles de carga y estrés a los que puede ser sometido el sistema	Rendimiento de los componentes de la arquitectura propuesta	Carga	(1-75 %) bien (75-90%) regular (90-100%) mal
			Estrés	(1-75 %) bien (75-90%) regular (90-100%) mal
Dependiente 2	Nivel de seguridad que posee el sistema	Niveles de seguridad de la arquitectura propuesta	Integridad de los Datos	Cantidad numérica
			Denegación de servicios	Cantidad numérica
			Transición insegura de HTTP a HTTPS	Cantidad numérica

Para el desarrollo de la presente investigación se utilizaron los siguientes métodos científicos:

Métodos Teóricos:

Analítico – Sintético: La investigación se encaminará a partir del análisis de los conceptos y métodos existentes relacionados al tema de hardware, seguridad y los componentes de los buscadores, con el fin de obtener resultados y así tenerlos en cuenta para el diseño de la arquitectura y configuraciones de seguridad. Se analizará la información y la documentación recopilada para la selección de las posibles propuestas, metodologías lógicas y adecuadas para el desarrollo de la arquitectura, para de esta forma realizar una síntesis de esta documentación.

Histórico – Lógico: La investigación se realizará a partir del estudio de la información existente de la problemática analizada. Teniendo en cuenta las investigaciones realizadas y los resultados obtenidos por otros autores, se establecerán similitudes entre la propuesta de solución y lo desarrollado anteriormente.

Inductivo – Deductivo: Este método estará orientado a establecer las generalizaciones sobre la base del estudio de casos particulares. A partir de la inducción de conocimientos, inferir deducciones que nos ayuden en el desarrollo del trabajo.

Modelación: Permitirá modelar los diagramas necesarios para obtener una idea más acertada acerca del análisis, diseño e implementación del sistema.

Métodos Empíricos:

Entrevista: Para la ejecución de este método se realizarán encuentros con los profesores, estudiantes y especialistas de diferentes áreas que poseen conocimientos acerca de la recuperación de información.

A raíz de la presente investigación **se espera como resultado**:

- El diseño de la arquitectura y configuraciones de seguridad para el buscador Orión v2.

La presente investigación se encuentra estructurada por un resumen, una introducción, 3 capítulos, y conclusiones, además de documentar bibliografías utilizados. Estos abordan, de manera completa, los aspectos fundamentales del desarrollo de la solución al problema planteado anteriormente.

En el capítulo 1: “Fundamentación teórica sobre diseño de arquitectura y configuraciones de seguridad en los sistemas de recuperación de información”, se presentan los elementos correspondientes a la fundamentación teórica de la propuesta, así como, los aspectos esenciales para el diseño de la arquitectura y la selección de la metodología de desarrollo y herramientas a utilizar.

En el capítulo 2: “Características y diseño de la arquitectura que se propone, requerimientos y análisis de la solución”, se abordan los componentes necesarios para el diseño de la arquitectura y los requerimientos con que debe cumplir. Además, se exponen las configuraciones de seguridad básicas que debe cumplir el sistema propuesto.

En el capítulo 3: “Validación para la propuesta de arquitectura y configuraciones de seguridad para el buscador Orión V2”, se realizan las pruebas necesarias para la validación de la propuesta. También se describe el criterio de experto analizado para la aceptación del diseño y las configuraciones de seguridad definidas en la propuesta como solución.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA SOBRE DISEÑO DE ARQUITECTURA Y CONFIGURACIONES DE SEGURIDAD EN LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

En el presente capítulo, se exponen aspectos teóricos concernientes a la investigación, además se abordan los principales conceptos referentes al tema. Se presenta un estudio de sistemas de búsqueda similares en el ámbito tanto nacional como internacional que brindaran datos y características a tener en cuenta para el análisis de la propuesta de solución. Además, se plasma la metodología, técnicas y herramientas que serán utilizadas en el desarrollo de la arquitectura de hardware y configuraciones de seguridad a utilizar para la propuesta de solución.

1.1 Conceptos asociados al dominio de la investigación

En el ámbito de la presente investigación es importante considerar que:

Recuperación de Información (RI)

Salton plantea que el área de RI es un campo relacionado con la estructura, análisis, organización, almacenamiento, búsqueda y recuperación de información (Salton, y otros, 1983), mientras que Abadal la define, como la disciplina que estudia la representación, la organización y el acceso eficiente a la información que se encuentra registrada en documentos (Abadal, y otros, 2005). Después de estudiados los conceptos anteriores, la autora considera que la RI no es más que la organización, almacenamiento, y acceso eficiente de la información.

Sistemas de Recuperación de Información (SRI)

Varios autores plantean que los sistemas de recuperación de información o buscadores son herramientas que permiten el acceso a la información alojada en la web y tienen como principal objetivo ayudar al usuario en el proceso de recuperar información con un alto valor y calidad acorde a su necesidad (Baeza-Yates, 1999). Un sistema de recuperación de información es definido como el proceso que trata la representación, almacenamiento, organización y acceso de elementos de información (Salton, y otros, 1983). Según los conceptos analizados acerca de los SRI, la autora considera que estos son herramientas que recolectan, almacenan y brindan información a los usuarios, en respuesta a las necesidades de los mismos.

Buscadores

Un concepto más amplio y asumido en la presente investigación es: un buscador es considerado un sistema o aplicación informática que permite la búsqueda de todo tipo de términos y palabras claves a partir del desarrollo de índices de archivos almacenados en servidores web, que opera indexando archivos y datos en la web (Definición ABC, 2015).

Un buscador emplea diferentes algoritmos y métodos para satisfacer la necesidad de información planteada por un usuario en una consulta en lenguaje natural especificada a través de un conjunto de palabras clave (Jaimes, y otros, 2005).

Una vez estudiados los conceptos analizados (Definición ABC, 2015) y (Jaimes, y otros, 2005), la autora asume que un buscador es una herramienta que permite satisfacer las necesidades de información de los usuarios, después de ejecutados los procesos de rastreo e indexación de los datos de la web.

Metabuscadores

A diferencia de los buscadores, un metabuscador no posee una base de datos propia, sino que utiliza la base de datos de estos para encontrar la información solicitada por el usuario. Su único trabajo consiste en combinar las mejores páginas que ha devuelto cada buscador, logrando así un mayor abanico de resultados con mucha mayor calidad (Consoft, 2002).

Directorios Temáticos

Algunos autores consideran los Directorios Temáticos Especializados (DTE) como una biblioteca digital que, en realidad, supera este propio concepto por tratarse colecciones de enlaces a recursos expertos, pero de alto valor añadido determinado por la selección y la evaluación del tema por expertos (Saz, 2003).

Los directorios temáticos especializados realizan una función de guía, puesto que localizan, seleccionan, organizan y comentan recursos de información de valor científico existentes en Internet. Por este motivo a lo largo del tiempo se han afianzado como herramientas fundamentales en el acercamiento a los usuarios de la información científica existente en la red (Tamayo, y otros, 2011).

Seguridad

La seguridad de la información es un proceso en el que se da cabida a un creciente número de elementos: aspectos tecnológicos, de gestión-organizacionales, de recursos humanos, de índole económico, de negocios, de tipo legal, de cumplimiento, etc.; abarcando no sólo aspectos informáticos y de telecomunicaciones sino también aspectos físicos, medioambientales, humanos, etc. (Bertolín, 2008).

En la presente investigación se entiende por configuraciones de seguridad en entornos de la RI y SRI, como el conjunto de medidas y técnicas de protección ante ataques que garantizan el correcto funcionamiento de la herramienta.

Arquitectura

La arquitectura de computadoras es el diseño conceptual y la estructura operacional fundamental de un sistema de computadora. Es decir, es un modelo y una descripción funcional de los requerimientos y las implementaciones de diseño para varias partes de una computadora, con especial interés en la forma en que la unidad central de proceso (CPU) trabaja internamente y accede a las direcciones de memoria (Vilca, 2017).

1.2. Estudio de sistemas homólogos

Para el estudio de los sistemas homólogos, se realiza un análisis de los SRI a nivel mundial; con el fin de comparar las características necesarias a tener en cuenta para el cumplimiento del objetivo general de la presente investigación.

1.2.1. Sistemas de Búsqueda Internacionales

En el ámbito internacional existe una gran gama de sistemas que se encargan de la búsqueda de información en la web. Muchos de estos sitios son privativos o simplemente no todos los usuarios tienen acceso a sus servicios. De manera general la información que brinda cada sistema de búsqueda acerca de su funcionamiento y estructura es muy desigual, por lo que a continuación se presentan algunas pautas a tener en cuenta para una mejor descripción de las características de los SRI estudiados:

- Información general
- Arquitectura de hardware
- Seguridad

- Popularidad del buscador (aceptación del público)

Google

- Según la M.Sc. Kryscia Daviana Ramírez Benavides, Google es el producto más notorio de una empresa norteamericana del mismo nombre. Fue fundado en 1998 por Larry Page y Serguei Brin, dos estudiantes de doctorado de la Universidad de Stanford. Google nace con la necesidad de obtener mejores respuestas de los motores de búsqueda.

  Google indexa más de 8 mil millones de páginas Web, aunque ofrecen más resultados gracias a los "rastreos profundos" (*crawlers*) como son:

  La búsqueda general (una vez al mes), que busca en la mayoría de la WWW.

  El *Fresh*, que rastrea en las páginas que se actualizan frecuentemente.

  La de noticias, que rastrea cada 10 minutos.

  Los servidores de recolección de datos navegan por Internet al estilo araña.

- Según una estimación entre los años 2000 -2006, de Google se tienen que:

  La granja de servidores de Google estaba compuesta por: 6000 procesadores, 12000 discos duros IDE (dos por maquina) en cuatro centros físicos: 2 en *Sillicon Valley* y 2 en Virginia. Cada centro tenía una conexión de fibra óptica de 2488 Mbit/s y otra de 622 Mbit/s.

  Más de 15.000 servidores con velocidades comprendidas entre el Intel Celeron de 533 MHz y el Pentium III a 1.4 GHz dual.

  Entre 200.000 y 450.000 servidores en esos años con uno o más discos duros de 80 GB por servidor y entre 2 y 4 GB de memoria RAM por máquina.

  En un estudio sobre el hardware de Google las cifras son exorbitantes, ronda un gasto de unos 250 millones de dólares en hardware: Entre 45.000 y 80.000 servidores, 69.000 las máquinas y 539 racks. (El rackes algo así como: 88 dualCPU 2Ghz Intel Xeon servers con 2Gbytes de RAM y un disco duro de 80Gbytes), (Benavides, 2015). Los racks están hechos a medida y pueden contener entre 40 y 80 servidores (cada rack tiene una conexión Ethernet a un router local que a su vez se conecta al router central utilizando una conexión de 1Gigabit).

  Actualmente, los servidores son ordenadores personales x86 que utilizan una versión personalizada de Linux. Se estima que se necesitan 20 MW para poder alimentar a 45.000 servidores, con coste asociado de unos 2 millones de dólares americanos al mes en facturas de electricidad.

- De las configuraciones de seguridad, no se encuentra información necesaria para el desarrollo de la investigación.
- Google se ubica entre los primeros lugares de las máquinas más rápidas del mundo, ya que tendría una capacidad de cálculo de 189 *teraflops*, un cálculo medio de 253 *teraflops* y si se redondea hacia arriba serían unos 316 *teraflops* de potencia.

  Google es el buscador más usado a nivel internacional, recibe cientos de millones de consultas cada día a través de sus diferentes servicios entre los que se encuentran 22 "características especiales" como son: clima, resultados deportivos, hora, calculadora, diccionario de búsqueda, mapas entre otras. Mantiene una interfaz amigable y es fácil de usar por los internautas. Está posicionado como el buscador más utilizado con un 62.74% de uso en el mercado en el 2015 (Suárez, 2016).

  A continuación, se muestra una imagen que contiene algunos de los servicios que posee la arquitectura de Google donde se presenta la relación entre los servidores y sus componentes.

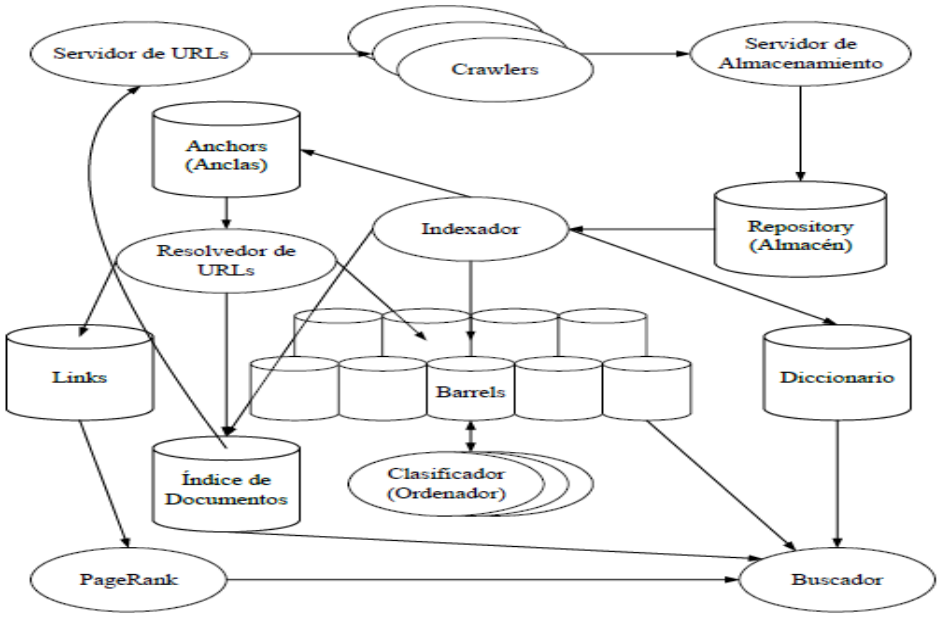

Ilustración 1: Arquitectura de google (Benavides, 2015)

Yahoo!

- El usuario puede personalizar sus búsquedas a través del servicio "*My Yahoo! Search*", de este modo, el internauta puede ir guardando así los enlaces que desee e incluso introducir comentarios personales.

  Este buscador cuenta con funcionalidades como:

Búsqueda rápida Yahoo! Que permite accede a las informaciones más populares (noticias, tiempo, mapas, etc.) directamente desde la caja de búsquedas, es decir las búsquedas guardadas por el usuario.

Yahoo! Search ofrece el resultado más relevante en relación con tu búsqueda.

Buscador Web Yahoo!, se encarga de responder a las consultas mostrando los resultados más pertinentes de la web.

Buscador de imágenes Yahoo!: Busca en más de 1,6 millones de imágenes de toda la red en consideración con la consulta realizada.

- Este buscador no ofrece información acerca de su estructura de hardware, por lo que no se puede analizar para la propuesta de solución.

- De las configuraciones de seguridad utilizadas por el SRI, no aparece información satisfactoria para el desarrollo de la investigación.

- El usuario puede personalizar sus búsquedas a través del servicio "*My Yahoo! Search*". De este modo, el internauta puede ir guardando así los enlaces que desee e incluso introducir comentarios personales. Además, el portal permite almacenar direcciones de sindicación de contenidos con el fin de rescatarlas y utilizarlas más tarde. Además, cuenta con numerosas herramientas útiles e interesantes (Yahoo!!, 2016).

  Este buscador se posiciona en el cuarto buscador más utilizado de la web, con un 7.79% de uso (Suárez, 2016).

Baidu

- Baidu es un motor de búsqueda en idioma chino. Su diseño es similar al de Google y ofrece 57 servicios de búsqueda y de comunidad, entre los que se encuentran: buscadores de música, imágenes, noticias, sitios web y muchos más. Quizá la característica más popular de Baidu, que otros motores de búsqueda como Google no ofrecen, es la posibilidad de efectuar búsquedas de archivos de audio (MP3, WMA/SWF...). Es usada fundamentalmente para la búsqueda de música pop china, y los resultados de la búsqueda son sorprendentemente precisos.
- A pesar de ser uno de los buscadores más utilizados, no se ofrece información acerca de su arquitectura de hardware, por lo que no se tiene en cuenta para el desarrollo de la propuesta de solución.

- Sobre las configuraciones de seguridad utilizadas por este buscador, no se encuentra información disponible.
- Baidu es uno de los buscadores más utilizados tanto en China, como es su extensión por Asia, por lo que se ha posicionado en el segundo lugar entre los SRI más visitados contando con 18.68% de popularidad (Suárez, 2016).

1.2.2. Buscadores Nacionales

A continuación, se presenta un estudio de los buscadores nacionales que permiten el flujo de información en el país, siendo de gran apoyo e importancia para su informatización. Para un mejor entendimiento acerca del estudio de estos SRI se realiza el análisis definido en el epígrafe 1.2.1.

Lupa

- Según la documentación que brinda el sitio lupa.upr.edu.cu es un buscador de contenidos en la red universitaria. Desarrollado en la Universidad de Pinar del Río por el Grupo de Desarrollo y Diseño Web de la Dirección de Informatización, con el propósito de localizar contenidos dispersos en sitios web y repositorios públicos y en sus enlaces transversales de alta velocidad (MINED, INFOMED, Cultura, Joven Club, UCI) (Lupa, 2017).
- Algunas de las funcionalidades que presenta el buscador son:

  Documentos similares: Esta se encarga de definir el tema de los documentos para que sean utilizados como punto de partida en la búsqueda.

  Filtrado de documentos por tipo de archivo: Lupa es capaz de organizar por tipo de archivo los resultados obtenidos en una consulta.

  Búsquedas relacionadas: Las búsquedas relacionadas están formadas por una nube de frases o términos representativos del conjunto de documentos devueltos en una consulta. Se utilizan para ayudar al usuario a direccionar su búsqueda hacia los resultados que desea.

  Carga única: Esta interfaz web funciona de forma autónoma del lado del cliente, por lo que solo es necesario cargarla una sola vez. Lo que facilita su empleo en redes lentas. Todas las operaciones de búsqueda, paginado y filtrado, se ejecutan intercambiando pequeños archivos con el servidor, potenciando así la rapidez de las mismas.

A continuación, se presentan los componentes pertenecientes al buscador:

1. ***Crawler* o robot:** programa que recorre la web y recuperan de forma automática los documentos encontrados
2. **Indexador:** recibe cada página o documento encontrado por el *crawler* y extrayendo la información, volcando esta representación en un índice de términos almacenado en una base de datos.
3. **Motor de búsqueda:** se encarga de analizar una consulta del usuario y buscar en el índice los documentos relacionados.
4. **Interfaz web:** permite al usuario escribir sus consultas y recibir una respuesta del buscador expresada como una lista de documentos relacionados.
5. Internet en busca de información en los populares buscadores internacionales.

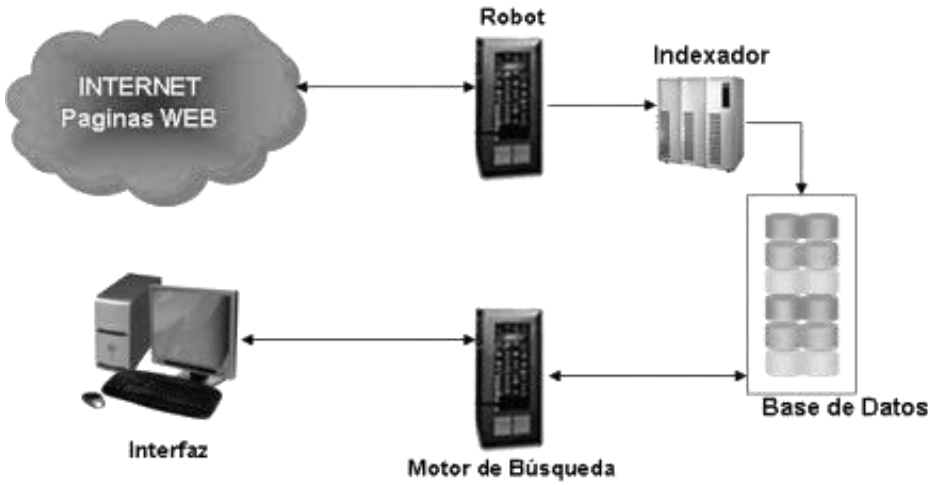

Ilustración 2: Arquitectura del buscador lupa (LUPA, 2017)

- El buscador no presenta se encuentra información acerca de la seguridad utilizada.

- El empleo de esta herramienta tiene un impacto directo en el aprovechamiento de los contenidos dentro del país que disminuye el tráfico de los usuarios de la red universitaria hacia internet en busca de información en los buscadores internacionales.

Orión

- Es un buscador cubano que se encuentra en desarrollo por la Universidad de Ciencias Informáticas. Surge en el 2011 con la idea de crear una herramienta de búsqueda de información con tecnologías libres, implementado en Linux. Según el departamento de desarrollo, este se sustenta por tres componentes fundamentales.

El componente de recolección o spider web es el encargado del rastreo y procesamiento de la información que se encuentra dispersa por toda la red. Como spider se utiliza el servidor Nutch**,** debido a las múltiples ventajas que ofrece a lo largo de todo el proceso de rastreo y almacenamiento de la información.

Nutch: *crawler* libre y de código abierto desarrollado en Java bajo la licencia de Apache. Este proporciona interfaces extensibles para implementaciones personalizadas. La arquitectura de Nutch es flexible y permite realizarle mejoras por parte de los usuarios a través de *plugins*.

Este es independiente del servidor de indexación lo que permite la integración con Solr (NutchWiki, 2015).

Para el componente de indexación se utiliza el servidor Solr, este es el encargado de recepcionar, procesar y almacenar toda la información rastreada por el spider (Cleverdon, 1997).

Solr: es un motor de búsqueda de código abierto basado en la biblioteca Java del proyecto Lucene, con APIs en XML/HTTP y JSON, sobresaltado de resultados, búsqueda por facetas, caché, y una interfaz para su administración (Apache Software Fundation, 2015).

Para el componente de visualización se utilizan los servidores web Nginx y para su desarrollo utiliza el marco de trabajo, Symfony, debido a la potencia que brinda en el desarrollo de sistemas web. Symfony**:** es un framework PHP construido con varios componentes independientes creados por el proyecto Symfony (Guiluz, 2013).

- En estos momentos la arquitectura del buscador cuenta dos servidores para el mecanismo de rastreo, tres servidores de indexación, un maestro y dos esclavos, dos servidores de visualización y un balanceador de carga para distribuir las peticiones de los usuarios en los servidores web, como se muestra en la (Ilustración 3).

- Acerca de las configuraciones de seguridad del buscador, en estos momentos no se encuentra funcionando, por lo que no se tiene en cuenta.

- El buscador no se encuentra disponible para el público en estos momentos por lo que no se posee información acerca de la evaluación de los usuarios hacia sus servicios.

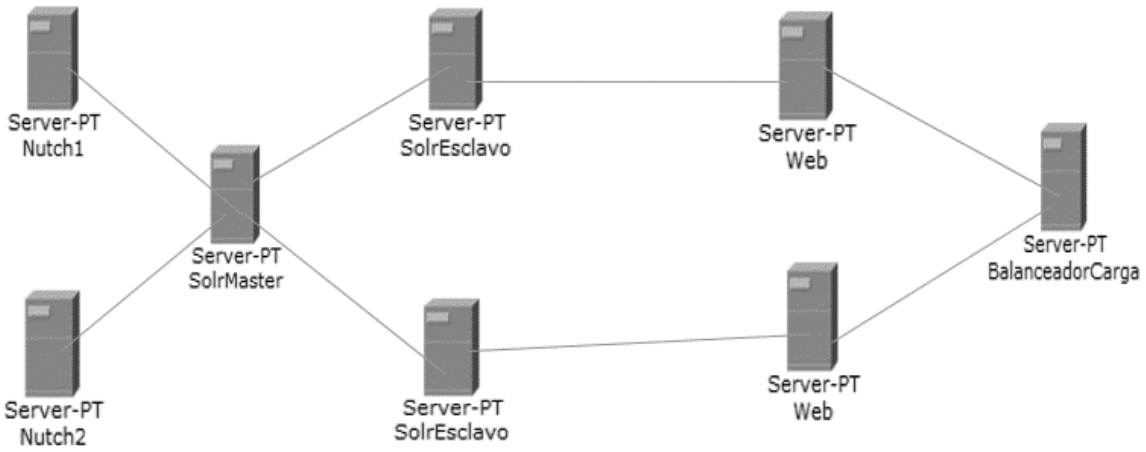

Ilustración 3: Arquitectura del buscador orión. fuente. (elaboración propia)

Plataforma de Servicios C.U.B.A

- Plataforma de servicios C.U.B.A, es un buscador que opera sobre el motor de búsqueda Orión. Esta plataforma forma parte de las acciones que se realizan para la informatización de la sociedad cubana. Esta permite recuperar información publicada solo bajo el dominino.cu, y es una alternativa para las personas que hoy solo tienen acceso a la intranet nacional.

- Este buscador posee una arquitectura de hardware donde se implantan 4 servidores para el componente de rastreo de información llamados Nutch. Estos se encargan de recorrer la web mediante una secuencia de procesos internos que garantizan la indización de cada uno de los enlaces a las páginas web y la extracción adecuada de los contenidos verdaderamente relevantes que serán consultados por las búsquedas de los usuarios.

  Para el proceso de indización se utilizan 4 servidores Solr, 2 maestros y 2 esclavos encargados de indexar los datos, para comunicar los servidores de rastreo con los de indexación, se utiliza un balanceador de carga que permite mejorar rendimiento del proceso de almacenamiento.

  La interfaz web está desplegada en 3 servidores que son accedidos mediante un balanceador de carga, que asigna las peticiones de los usuarios a un servidor web especifico y la conexión con los servidores de indexación se establece mediante 2 balanceadores de carga destinados a establecer un balanceo de las peticiones ejecutadas por los usuarios (Ilustración 5).

- No se encontró mucha información disponible, por lo que se analizó solo las configuraciones básicas referente al uso de cortafuegos como:

  Nutch: Permitir solo el tráfico saliente hacia internet.

  Solr maestros: Permitir solo el tráfico entrante de los servidores de balanceo de carga que se encuentran entre los nutch y los sorl.

  Solr esclavos: Permitir solo de permitir el tráfico entrante de los solr maestros, los balanceadores de carga (Solr-Web).

  Web: Solo permitir el tráfico entrante de consultas al buscador a través de los servidores de balaceo de carga que se encuentran entre los usuarios y la interfaz.

  Balanceadores de carga (Nutch-Solr): Permitir solo el tráfico entrante de los servidores de nutch.

  Balanceadores de carga (Solr-Web): Permitir solo el tráfico entrante de los servidores web.

  Balanceadores de carga (Usuarios-Web): Permitir solo el tráfico entrante de consultas de usuarios por el protocolo HTTPS y el puerto 8080.

- El buscador posee una buena aceptación entre los usuarios nacionales ya que se encuentra al alcance de toda la población. A continuación, se presenta las visitas realizadas al sitio en un día, demostrando el flujo de usuarios que acceden al mismo, aunque no cuenta con gran número de usuarios. Registran un usuario cada 59 segundos como promedio, además de mostrar el horario con más concurrencia, con un promedio de entre las 10am y las 2 pm.

A continuación, se muestra el diseño de la arquitectura de hardware de la plataforma:

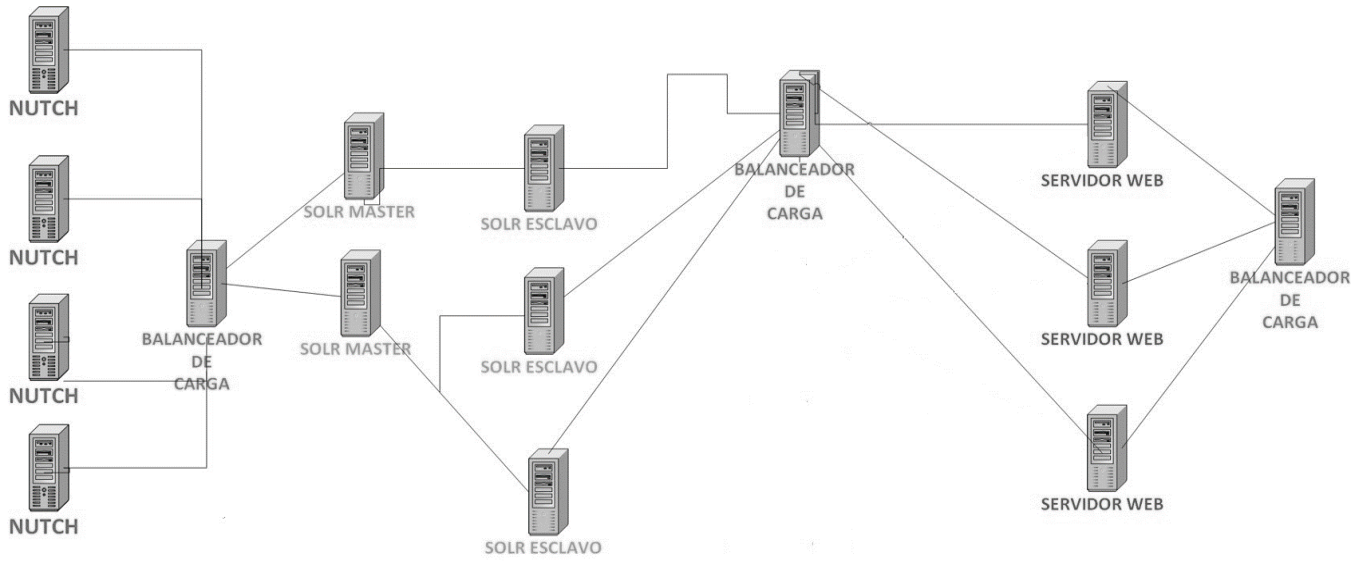

Ilustración 4:Arquitectura de la Plataforma de Servicios C.U.B.A (CIDI, 2017)

Resultados del estudio de los sistemas similares existentes

Luego de realizar un estudio a los motores de búsqueda, tanto en el ámbito nacional como internacional, se llegó a las siguientes conclusiones:

- Los buscadores internacionales no brindan mucha información acerca de las arquitecturas de hardware y las configuraciones de seguridad que utilizan.
- La arquitectura base de los SRI estudiados es similar y cuenta con tres mecanismos fundamentales: rastreo, indexación y visualización.
- Se decide utilizar como base la arquitectura propuesta por La Plataforma de Servicios C.U.B.A referente a los mecanismos de rastreo, indexación y visualización ya que cumple con las necesidades fundamentales para lograr un correcto rendimiento en estos mecanismos en el buscador Orión v2.
- Es necesario diseñar una nueva arquitectura para los componentes restantes del buscador.

1.3. Análisis del entorno de desarrollo de la investigación

A continuación, se realiza una descripción de las tecnologías, herramientas y metodología utilizada para el desarrollo de la propuesta de solución para la arquitectura y configuraciones de seguridad en los sistemas de búsqueda, específicamente el buscador Orión v2.

1.3.1. Metodología de desarrollo

Una metodología de desarrollo, en ingeniería de software, es un conjunto de herramientas, técnicas, procedimientos y soporte documental encaminados a estructurar, planificar y controlar el proceso de desarrollo de forma organizada y lógica, que tiene como objetivo apoyar a los desarrolladores en la creación de un nuevo software (Zambrano, 2015).

Como metodología de desarrollo se selecciona (AUP-UCI), versión modificada por la universidad, de forma tal que se adapte al ciclo de vida definido para la actividad productiva de los centros de desarrollo de la UCI. Esta se basa en las fases que propone la metodología AUP (Inicio, Elaboración, Construcción, Transición) ya que mantiene la fase de inicio, pero modificando el objetivo de la misma, y se unifican las otras tres fases en una sola, a la cual se le denomina Ejecución y se decide adicionar una nueva fase llamada Cierre (Tabla 2) (Sánchez, 2015).

Esta metodología utiliza las buenas prácticas que propone el Modelo CMMI-DEV v1.3. Este constituye una guía para aplicar las mejores prácticas en una entidad desarrolladora. Estas prácticas se centran en el desarrollo de productos y servicios de calidad (Sánchez, 2015).

Tabla 2: Ciclo de vida de la metodología AUP-UCI

Fases AUP	Fases Variación AUP-UCI	Objetivos de las fases (Variación AUP-UCI)
Inicio	Inicio	En esta fase se realiza un estudio inicial de la organización cliente que permite obtener información fundamental acerca del alcance del proyecto, realizar estimaciones de tiempo, esfuerzo y costo y decidir si se ejecuta o no el proyecto.
Elaboración	Ejecución	En esta fase se ejecutan las actividades requeridas para desarrollar el software, incluyendo el ajuste de los planes del proyecto considerando los requisitos y la arquitectura. Durante el desarrollo se modela el negocio, obtienen los requisitos, se elaboran la arquitectura y
Construcción		
Transición		

		el diseño, se implementa y se libera el producto.
	Cierre	En esta fase se analizan tanto los resultados del proyecto como su ejecución y se realizan las actividades formales de cierre del proyecto.

1.3.2. Herramientas

Herramienta Case Visual Paradigm 8.0

Es una herramienta CASE multiplataforma, que soporta el ciclo completo de desarrollo de software: análisis, diseño, implementación y pruebas. Facilita la construcción de aplicaciones informáticas con un menor coste que destacan por su alta calidad y contribuye a mejorar la experiencia de usuario mediante el diseño de un gran número de artefactos de ingeniería de software. Permite la generación de bases de datos, conversión de diagramas entidad-relación a tablas de base de datos, mapeos de objetos y relaciones, ingeniería directa e inversa, la gestión de requisitos de software y la modelación de procesos del negocio (Visual Paradigm, 2014).

Herramienta propuesta para el modelado de arquitectura ya que esta brinda la posibilidad de documentar todo el trabajo sin necesidad de utilizar herramientas externas, también genera documentación en formatos HTML y PDF. Además de estar disponible en múltiples plataformas: Microsoft Windows (98, 2000, XP, o Vista), Linux, Mac OS X, Solaris o Java.

Visio Profesional

Es uno de los programas informáticos, exclusivamente para Windows, más versátiles y comúnmente utilizados en el desarrollo de propuestas de diseño web. Ofrece una librería formada por una cincuentena de elementos gráficos, que puede ser enriquecida con elementos externos, y aporta una caja de búsqueda para facilitar el acceso al elemento a partir de su nombre.
Visio permite la exportación a pdf, html, sgv, tiff, jpeg, gif, png y visio; e importa html, sgv, tiff, jpeg, gif, png y visio. Soporta la inclusión de anotaciones y notas a pie de página, edición colaborativa de prototipado y la creación de prototipos dinámicos (Gutiérrez, y otros, 2010).

- Simplifica y comunica información compleja con diagramas vinculados a los datos que puedes crear en tan solo unos clics.

- Ayuda a trabajar visualmente, ya sea capturando rápidamente un diagrama de flujo que surgió a raíz de una lluvia de ideas en una pizarra, creando un organigrama, documentando un proceso empresarial o dibujando un plano de planta.
- Crear diagramas profesionales rápidamente con un gran conjunto de diagramas pre-diseñados.
- Crear diagramas de flujo de datos, diagramas de flujo de programas, y más, que permiten iniciar al usuario en los lenguajes de programación.

Apache Solr

Solr es la plataforma de búsqueda de código abierto del proyecto Apache *Lucene*. Sus características principales incluyen potentes búsquedas sobre texto completo, marcado de coincidencias, búsqueda por facetas, *clustering* dinámico, integración de bases de datos y manejo de documentos avanzados (por ejemplo, Word, PDF). Además, Solr es altamente escalable, proporcionando búsquedas distribuidas y replicación de índices. Por ello, proporciona las funciones de navegación y búsqueda de muchos de los sitios más grandes del mundo de Internet. Solr utiliza la librería de *Lucene* como núcleo para la indexación y búsqueda de texto completo (Fonollosa, y otros, 2012). Este componente ofrece múltiples ventajas como:

Ventajas de uso de Solr (Leyva, y otros, 2016):

- Capacidades avanzadas de búsqueda a texto completo.
- Optimizado para elevados volúmenes de tráfico.
- Interfaces abiertas basadas en estándares abiertos (XML, JSON, HTTP).
- Flexibilidad y adaptabilidad a través de extensas opciones de configuración.
- Caché altamente configurable.
- Extracción de metadatos.
- Soporte para varios núcleos.
- Sobresaltado de los resultados.
- Auto-sugerencias para completar las consultas de los usuarios (Apache Software Fundation, 2015).
- Escalabilidad – Replicación eficiente hacia otros Servidores de Búsqueda de Solr (Seta, 2010).

Apache Nutch

La arquitectura del buscador Orión cuenta en su estructura con el componente de recolección o spider web. Para su uso se utiliza Nutch, debido a las múltiples ventajas que ofrece a lo largo de todo el proceso de rastreo y almacenamiento de la información (NutchWiki, 2015).

Algunas ventajas del uso de Nutch (Leyva, y otros, 2016):

- Licencia Apache Software Foundation (ASF).
- Recolección de información en formato PPT, DOC, PDF, HTML, XML, TXT, RTF, GIF, JPG, PNG, entre otros.
- Arquitectura distribuida.
- Extensibilidad.
- Multiplataforma.
- Lenguaje Java.
- Amplia y fuerte comunidad de desarrollo (Apache).
- Perspectivas de desarrollo.
- Documentación (Idioma inglés).

Nutch es una implementación de código abierto en Java de un *crawler,* proporcionando todas las herramientas que necesita para ejecutarlo. Entre sus bondades pueden destacarse la transparencia y el entendimiento del sistema, ya que al ser de código abierto se puede ver cómo funcionan sus algoritmos. Otro de sus puntos fuertes es la extensibilidad, Nutch es muy flexible y permite que los desarrolladores puedan añadir funciones de filtrado de recursos, de indexación o de procesamiento para nuevos tipos de recursos. La principal ventaja de Nutch es que está basado en lo referente a la generación de índices en las librerías del proyecto *Lucene.* Por lo que, con la configuración adecuada, puede generar índices compatibles con el motor de búsqueda *Lucene* y en consecuencia con Solr. Nutch habitualmente puede funcionar a una de estas tres escalas: sistema de archivos local, intranet, o en la web entera. (Fonollosa, y otros, 2012).

A continuación, se muestra un estudio de los posibles servidores webs a utilizar en la propuesta de solución:

Apache

Apache es un poderoso servidor web, cuyo nombre proviene de la frase inglesa "*a patchy server*" y es completamente libre. Una de las ventajas de Apache, es que es un servidor web multiplataforma, es decir, puede trabajar con diferentes sistemas operativos y mantener su excelente rendimiento. Este estupendo servidor es utilizado principalmente, para realizar servicio a páginas web, ya sean estáticas o dinámicas y una de sus principales características es que se integra a la perfección con otras aplicaciones (Balkhi, 2013).

Este servidor web de código abierto implementa el protocolo HTTP 1.1, caracterizado fundamentalmente por su alto nivel de configuración, modularidad, robustez y estabilidad. Está desarrollado bajo la licencia ASF por *The Apache Software Foundation* (Netcraft, 2013).

Entre sus características podemos destacar:

- Multiplataforma.
- Estructura modular y API de desarrollo de módulos para hacerlo extensible, lo que permite aumentar sus funcionalidades mediante la adicción de complementos. Para ello proporciona una librería de funciones (kit de desarrollo) que simplifica el proceso de creación de extensiones.
- Soporte. Debido a su gran divulgación a lo largo de los años, pueden encontrarse gran cantidad de documentación en línea acerca de cualquiera de sus componentes, además de una gran cantidad de usuarios con conocimientos técnicos, foros especializados, libros, etc.
- Código abierto. Se distribuye bajo licencia Apache 2.0, dando la libertad al usuario para modificar, distribuir y vender las modificaciones que se hagan (con pequeñas restricciones como informar que el nuevo desarrollo es un proyecto basado en el código de Apache y adjuntando el aviso de licencia original).
- Utiliza la última versión estandarizada del protocolo HTTP por lo que asegura la compatibilidad y el funcionamiento con todas las aplicaciones que requieran su uso.
- Configuración sencilla gracias a la sintaxis y el formato de los ficheros de texto.

Nginx

Nginx es un servidor web y proxy libre, de código abierto y de alto rendimiento, ofrece estabilidad, un gran conjunto de características básicas de los servidores web, configuración sencilla, y bajo consumo de recursos (Nginx, 2015). Este se integra con diferentes tecnologías web como Apache y PHP y presenta soporte de una amplia y activa comunidad (Kholodkov, 2015).

Haciendo uso de s*ockets* asíncronos utiliza un proceso por núcleo para manejar miles de conexiones, lo que permite un consumo de carga de la CPU y la memoria mucho más ligera (Nedelcu, 2010).

Algunas ventajas de Nginx (Molina, 2015):

- Los ficheros de configuración son simples archivos de texto divididos en bloques de directivas, con campos de valores que pueden ser cambiados por el administrador.
- Flexibilidad.
- Potente.

- Proxy inverso liviano de alto rendimiento.
- Bajo consumo de recursos.

Después de un análisis sobre los servidores estudiados para el componente de visualización, se seleccionó NGINX en su versión 1.10 como servidor web y de aplicaciones por sus principales características que se asocian con el rendimiento, la escalabilidad y la eficiencia de costes. Es mucho más ligero que Apache, por lo tanto, realiza menos funciones. Es un servidor de archivos estáticos, índices y auto indexado, permitiendo el balance de carga y la tolerancia a fallos.

Acunetix Web Vulnerability Scanner

Es la herramienta utilizada en la Universidad para la detección de vulnerabilidades en sitios y aplicaciones web. Esta herramienta posee un componente que facilita la realización de pruebas a formularios y a áreas protegidas por contraseña. Permite realizar varias peticiones simultáneamente, siendo capaz de explorar cientos de páginas sin interrupciones. Para la realización de las pruebas de seguridad al subsistema se utiliza esta herramienta en su versión 9.5 (Ortega, 2014).

Conclusiones del Capítulo

Con el desarrollo del capítulo quedan trazados los fundamentos teóricos, las herramientas, conceptos, y metodología, necesarios para la ejecución de la presente investigación. Mediante el análisis de los sistemas semejantes de RI se comparan los elementos necesarios para las posibles vías de solución del problema planteado.

La fundamentación abordada en el capítulo, permitió un mejor desempeño y desarrollo de la investigación, definiéndose así las siguientes conclusiones:

- En los antecedentes de los SRI analizados en el ámbito nacional e internacional, se evidenció la existencia de motores de búsqueda muy destacados como son: Google, La Plataforma de servicios C.U.B.A, Orión V1, Lupa. Estos utilizan arquitecturas similares y cuentan con los mecanismos fundamentales, de todo buscador, rastreo, indexación y visualización.

- El estudio de las herramientas, conceptos y tecnologías permitió definir la base tecnológica necesaria para dar cumplimiento con la propuesta de arquitectura y configuraciones de seguridad para el buscador Orión v2.

- Se definen las herramientas Case Visual Paradigm 8.0 y Visio Profesional 2016 para el modelado de la propuesta y los servidores Solr, Nutch, y Nginx para el proceso de desarrollo y simulación de la arquitectura del buscador Orión v2.

CAPÍTULO 2: CARACTERÍSTICAS Y DISEÑO DE LA ARQUITECTURA QUE SE PROPONE, REQUERIMIENTOS Y ANÁLISIS DE LA SOLUCIÓN

2.1 Introducción

En este capítulo se realiza una descripción de la propuesta de solución, teniendo en cuenta el diseño de la arquitectura de hardware y las configuraciones de seguridad para el buscador Orión v2. También se presentan los requerimientos con los que debe cumplir la arquitectura definida. Además, se seleccionan el estilo arquitectónico y los patrones de diseño necesarios, evidenciando su utilización en la propuesta diseñada.

2.2 Descripción de la propuesta de solución

Dadas las necesidades planteadas en la situación problemática descrita en el capítulo uno del presente trabajo de diploma, la solución propuesta constituye un diseño de arquitectura de hardware y configuraciones de seguridad para el buscador Orión v2. La misma permite la estructuración de los componentes necesarios para proporcionar un mejor rendimiento con el menor costo posible, teniendo en cuenta las necesidades del proyecto. Además, se basa en los siguientes principios:

Disponibilidad: La disponibilidad es la cualidad de un sistema para mantenerse operativo principalmente ante contingencias: “Cada petición eventualmente recibe una respuesta”. La Alta Disponibilidad (HA) desde la visión de Eric Brewer implica mantener el servicio funcionando ante (Antiñanco, 2014):

Caídas y fallas de discos
Actualizaciones de bases de datos
Actualizaciones de software
Actualizaciones de sistema operativo
Cortes de energía
Cortes en la red
Movimiento físico del equipo
La propuesta de solución garantiza el concepto de disponibilidad mediante la creación de servidores de salva de la información y mecanismos de replicación.

Integridad: La integridad es la garantía de que nadie pueda acceder a la información ni modificarla sin contar con la autorización necesaria (Salta, 2015). La verificación de la integridad de los datos consiste en determinar si se han alterado los datos durante la transmisión (accidental o intencionalmente). En el caso de la propuesta de solución el concepto de integridad se garantiza mediante la restricción física del acceso a los servidores de almacenamiento de datos a personal no autorizado y la creación de mecanismos para que los usuarios del buscador solo accedan a la información permitida según sus roles.

Respaldo: El respaldo de datos se basa en efectuar una copia de todos o algunos archivos que se encuentran en el medio de almacenamiento de una o varias computadoras, servidores o en otros medios diferentes, para poder recuperarlos en otro momento si se pierden o se dañan los archivos originales. Este proceso debe ser fiable, eficaz y robusto, debido a que un error durante el proceso puede provocar que los datos se deterioren (Reinoso, y otros, 2007).

En la propuesta se evidencia el respaldo de datos a través de la réplica de la información, garantizada mediante la distribución de información entre servidores.

Interoperabilidad: La capacidad que tienen dos o más sistemas o componentes de intercambiar información y usar esa información que se ha intercambiado (Tamayo, y otros, 2011).

La arquitectura permite la integración entre componentes, una de las características del estilo arquitectónico utilizado.

Balanceo de carga: Clúster que permite que un conjunto de servidores comparta la carga del trabajo y del tráfico a sus clientes. Está compuesto por uno o más ordenadores (llamados nodos) que actúan como *front-end* del clúster y se ocupa de repartir las peticiones de servicio que reciba el clúster a los otros ordenadores que forman su *back-end* (Sinisterra, y otros, 2012).

Los balanceadores son utilizados para distribuir la información entre otros servidores, facilitando el funcionamiento y posibles sobrecargas de los mismos. En el caso de la propuesta de solución el concepto de balanceo se garantiza mediante la creación de servidores de balanceo de carga que permiten disminuir la carga de procesamiento de los servidores de indexación, visualización y datos de usuarios.

Seguridad: La seguridad de la información es la encargada de garantizar las medidas de protección para el desarrollo y sostenibilidad de la actividad de negocio (Bertolín, 2008).

La seguridad es utilizada para la protección de datos e información, tanto la almacenada en los servidores, como la que se encuentra transitando por la web. Esto se garantiza con la utilización de cortafuegos y restringiendo el flujo de datos al necesario solamente.

Escalabilidad: La escalabilidad indica la habilidad de un sistema para reaccionar y adaptarse sin perder calidad, o bien manejar el crecimiento continuo de trabajo de manera fluida, o bien estar preparado para hacerse más grande sin perder calidad en los servicios ofrecidos (Antiñanco, 2014). La escalabilidad del sistema permitirá que se puedan modificar, o adaptarse a nuevos cambios sin poner en riesgo el propio sistema. En el caso de la propuesta de solución el concepto de escalabilidad se garantiza mediante la creación de una arquitectura distribuida en cada uno de los componentes principales del buscador:

- componente de rastreo
- componente de indexación
- componente de visualización

2.3 Estado actual

Para una mejor comprensión del entorno donde se evidencia el problema en cuestión, en el presente trabajo de diploma se decide realizar un modelo que defina el estado actual que presenta el buscador. El objetivo del modelado es comprender y describir los componentes más importantes dentro del estado actual del sistema.

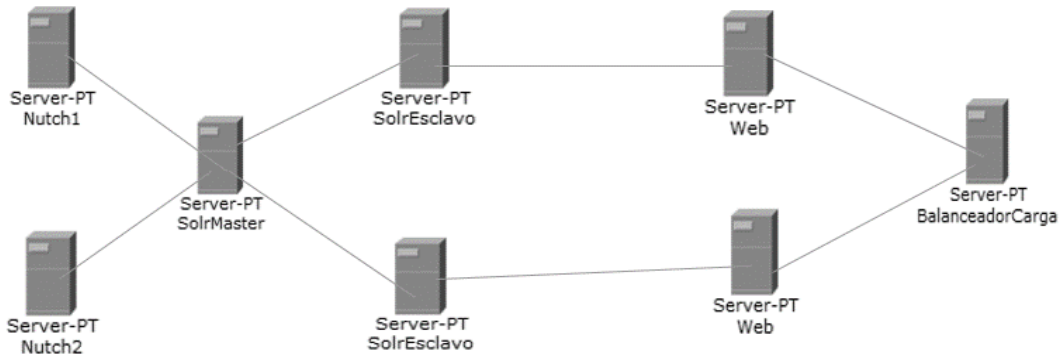

Ilustración 5:: Arquitectura de Hardware del Buscador Orión. Fuente (Elaboración Propia)

Descripción de los componentes de la Arquitectura actual de Orión:

La arquitectura cuenta con 2 servidores para el componente de rastreo de la información. Estos se encargan del recorrido por la web y la extracción adecuada de los contenidos que serán consultados por las búsquedas de los usuarios. Además, para el componente de indexación se cuenta con 3 servidores, 1 solr maestro y 2 esclavos encargados de la indexación de los datos. Para el componente de visualización se tienen 2 servidores web que son accedidos mediante dos balanceadores de carga, que asigna las peticiones de los usuarios a un servidor web específico.

2.4 Requerimientos

Para una correcta definición de la arquitectura se definen una serie de requisitos que la aplicación debe cumplir. La Ingeniería de Requerimientos ayuda a los ingenieros de software a entender mejor el problema en cuya solución trabajarán. Incluye el conjunto de tareas que conducen a comprender cuál será el impacto del software sobre el negocio, qué es lo que el cliente quiere y cómo interactuarán los usuarios finales con el software" (Pressman, 2006).

El levantamiento de requerimientos es la actividad que permite definir las necesidades del negocio, así como las funcionalidades a implementar. Un correcto levantamiento de requerimientos permitirá posteriormente realizar un correcto levantamiento de las funcionalidades con las que debe contar la arquitectura.

A partir del estudio del problema y la entrevista realizada al grupo de desarrollo se identificaron los siguientes requerimientos:

Requerimientos:

Gestionar hardware para el componente de rastreo.

Gestionar hardware para el componente de indexación.

Gestionar hardware para el componente de visualización.

Gestionar hardware para el componente de procesamiento.

Gestionar hardware para el componente de estadística.

Gestionar hardware para el componente de datos de usuarios.

Gestionar hardware para el componente de apoyo a la toma de decisiones.

Gestionar configuraciones de seguridad para el componente de rastreo.

Gestionar configuraciones de seguridad para el componente de indexación.

Gestionar configuraciones de seguridad para el componente de visualización.

Gestionar configuraciones de seguridad para el componente de procesamiento.

Gestionar configuraciones de seguridad para el componente de estadística.

Gestionar configuraciones de seguridad para el componente de datos de usuarios.

Gestionar configuraciones de seguridad para el componente de apoyo a la toma de decisiones.

Gestionar balanceo, alta disponibilidad y replicación para el componente de rastreo.

Gestionar balanceo, alta disponibilidad y replicación para el componente de Indexación.

Gestionar balanceo, alta disponibilidad y replicación para el componente de visualización.

Gestionar balanceo, alta disponibilidad y replicación para el componente de procesamiento.

Gestionar balanceo, alta disponibilidad para el componente de datos estadísticos.

Gestionar balanceo, alta disponibilidad para el componente de datos de usuarios.

Gestionar balanceo, alta disponibilidad para el componente de apoyo a la toma de decisiones.

Proteger los datos basándose en la confidencialidad, integridad y disponibilidad.

Gestionar acceso a los componentes de rastreo, indexación, estadística, procesamiento, visualización, datos de usuarios, y apoyo a la toma de decisiones.

Gestión de trazas en los componentes de rastreo, indexación, estadística, procesamiento, visualización, datos de usuarios, y apoyo a la toma de decisiones.

2.5 Propuesta del diseño de arquitectura de hardware para el buscador Orión v2

La arquitectura propuesta para el despliegue del buscador Orión v2, está estructurada por componentes, la relación que se establece entre ellos se expresa a continuación:

- Se presentan 4 servidores nutch pertenecientes al componente de rastreo distribuido en las categorías de deporte, cultura, noticias, e imágenes que son los encargados de buscar información en la web a través de los *spiders*.

Se propone como política de seguridad para estos servidores la utilización de cortafuegos, donde solo se permita el tráfico saliente de los *spiders* hacia internet para la búsqueda de información.

- A estos se conectan dos balanceadores de carga (Nutch-Solr), que distribuyen la información rastreada hacia 2 servidores solr maestros, estos servidores solo permiten el tráfico entrante desde los balanceadores (Nutch-Solr). Los servidores solr maestros replican la información recibida a 4 servidores sorl esclavos, todos pertenecientes al componente de indexación encargados del indexado de documentos en la arquitectura. Dichos servidores solo permiten el tráfico entrante de los balanceadores (Solr-Web) y los solr maestros.

- El servidor de procesamiento se conecta a los 4 servidores sorl esclavos y a 2 de datos de usuarios, permitiendo solo el tráfico entrante de información de estos, hacia él.

- En el caso del servidor de datos estadísticos se conecta al servidor de procesamiento para la recopilación de información necesaria para el análisis estadístico. Este servidor estadístico solo permite el tráfico entrante del servidor de procesamiento.

- El servidor de apoyo a la toma de decisiones se relaciona con el servidor de datos estadísticos con el objetivo de consultar la búsqueda de información útil para el perfeccionamiento y la toma de decisiones en el funcionamiento del buscador.

- En el componente de visualización se cuenta con 2 servidores de balanceo de carga (Usuarios-Web) q permiten el enlace entre los usuarios y los demás servidores, permitiendo solo en tráfico entrante de consultas de usuarios por el protocolo HTTPS. Dichos balanceadores se conectan a 3 servidores web, encargados de recibir y devolver la información demandada por el usuario. Estos nodos solo permiten el tráfico entrante de los balanceadores, analizados anteriormente.

- Entre los servidores web y los servidores sorl esclavos, se conectan 2 balanceadores (Solr-Web), que permiten solo el tráfico entrante de información proveniente de los servidores web.

- En el caso del balanceador (DatosUsuarios-Web) reciben consultas de los servidores web, y la distribuyen hacia los nodos de datos de usuarios, permitiendo solo el tráfico entrante de los servidores web.

2.5.1 La arquitectura se organiza de la siguiente manera

Componente de rastreo (Ilustración 6): la arquitectura de este mecanismo puede ser muy variable, en este caso se propone organizarla teniendo en cuenta el tamaño y las categorías de la red que se desea rastrear. Si la red posee un tamaño muy grande entonces un número reducido de *spìders* o servidores de rastreo tardarían mucho tiempo en realizar un proceso de rastreo completo o en llevar a cabo un proceso de actualización de contenidos con la calidad requerida. En el caso de la ilustración 6, se quiere realizar todo el rastreo de una red que posee contenidos de las categorías deporte, cultura, noticia e imágenes por lo que se propone el uso de 4 servidores de rastreo, la periodicidad de este proceso depende de cuán rápido actualizan la información los sitios.

El uso de un balanceador de carga en el proceso de rastreo e indexación, permite no sobrecargar los servidores que se encargan de recibir los datos rastreados e indexarlos, otra razón para su uso es que si se decide almacenar la información en servidores organizados por categorías entonces este balanceador sería el encargado de direccionar la inserción de datos en cada servidor correspondiente.

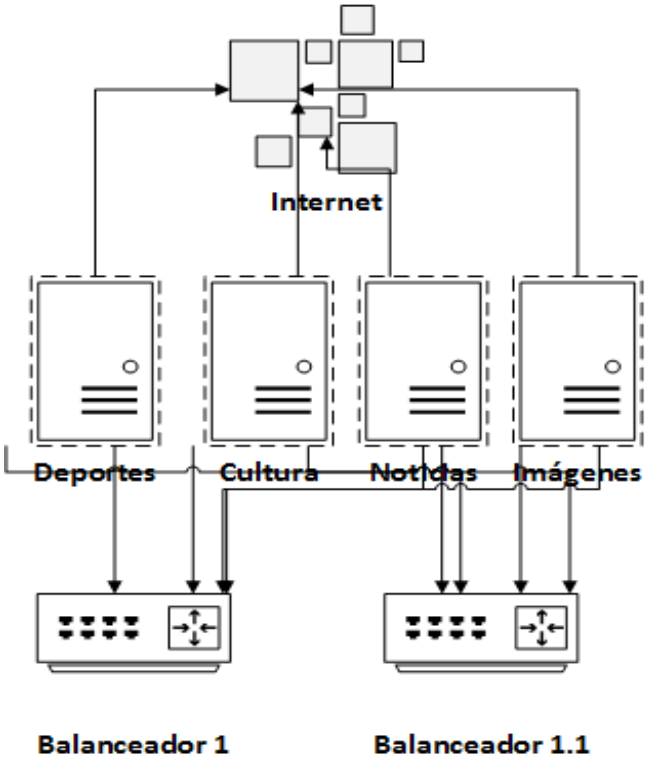

Ilustración 6:Componente de Rastreo

La propuesta de hardware para cada servidor es la siguiente:

Los servidores nutch: mínimo de 8GB de RAM.

Los servidores nutch: mínimo 500GB de almacenamiento.

Los servidores nutch: mínimo un procesador de 4 núcleos.

Para los balanceadores de carga (Nutch-Solr): mínimo 4 GB de RAM.

Para los balanceadores de carga (Nutch-Solr): mínimo 50 GB de almacenamiento.

Para los balanceadores de carga (Nutch-Solr): mínimo un procesador de 2 núcleos.

Componente de indexación (Ilustración 7): la arquitectura de este mecanismo depende del volumen de los datos que se van a almacenar, por esta razón es aconsejable realizar un estudio del tamaño de la web a la que se pretende acceder y definir entonces un número de servidores adecuado para almacenar estos datos.

Este componente se encarga del indexado de documentos, marcado de coincidencias, integración de bases de datos y manejo de documentos avanzados (por ejemplo, Word, PDF). Para la misma se propone un esquema básico de 6 servidores, de ellos dos servidores maestros encargados de recibir todo el flujo de inserción de datos, almacenarlos y llevar a cabo un proceso de réplica de la información en los 4 servidores esclavos.

La propuesta de hardware para cada servidor es la siguiente:

Los servidores solr maestro: mínimo de 8GB de RAM.

Se recomienda que cada servidor solr maestro: mínimo 500GB de almacenamiento.

El servidor solr maestro: mínimo un procesador de 4 núcleos.

Los servidores solr esclavos: mínimo 8GB de RAM a los esclavos.

Los servidores solr esclavos: mínimo 500GB de almacenamiento.

El servidor solr esclavos: mínimo un procesador de 4 núcleos.

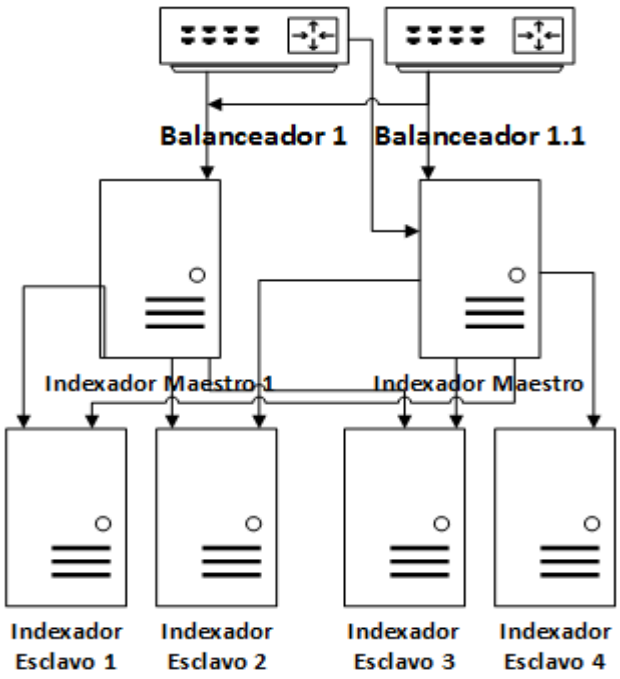

Ilustración 7:Componente de Indexación

Componente de procesamiento (Ilustración 8): Este componente es el encargado de ejecutar los procesos de categorización de documentos y la definición de perfiles de usuarios utilizando técnicas de minería de texto, la estructura que se propone, cuenta con un servidor de altas prestaciones para el procesamiento de dos tipos de datos:

- Documentos extraídos de los servidores de indexación.
- Consultas insertadas por cada uno de los usuarios, almacenadas en los servidores de datos de usuarios.

La propuesta de hardware para el servidor de procesamiento y el de datos de usuarios es la siguiente:

El servidor de procesamiento: mínimo 16 GB de RAM.

El servidor de procesamiento: mínimo 500 GB de almacenamiento.

El servidor de procesamiento: mínimo un procesador de 7 núcleos.

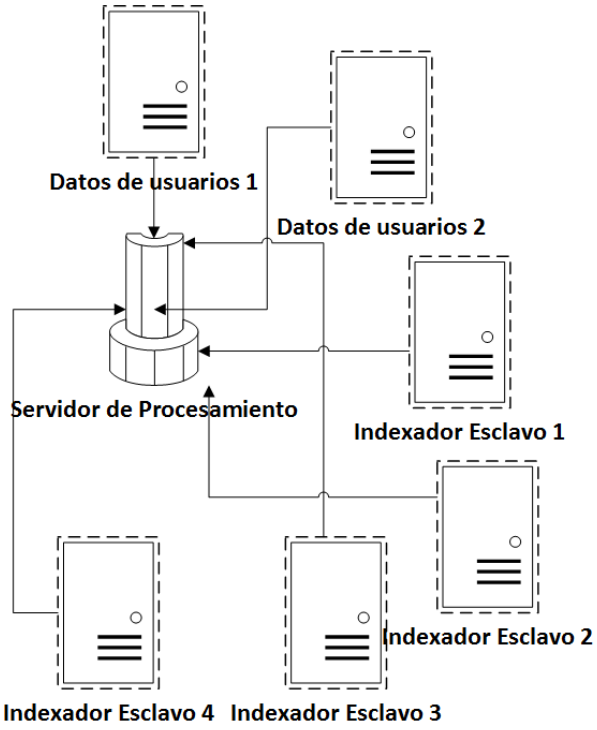

Ilustración 8:Componente de Procesamiento

Componente de visualización (Ilustración 9): Este componente es el encargado de brindar a los usuarios la interfaz necesaria para la inserción de consultas y la visualización de resultados. La arquitectura, está compuesta por 3 servidores web, 2 servidores de datos de usuarios y 3 balanceadores de carga. Dos de los cuales se encuentran entre los usuarios y la aplicación, el primero con el objetivo de recibir y distribuir las peticiones de los usuarios entre tres servidores de interfaces web, y el segundo para recibir esta tarea en caso de que el otro falle y así lograr un mejor rendimiento. Mientras que el tercer balanceador de carga es para distribuir la carga de inserción de datos de usuarios en dos servidores, entre los que se realiza un proceso de réplica de datos.

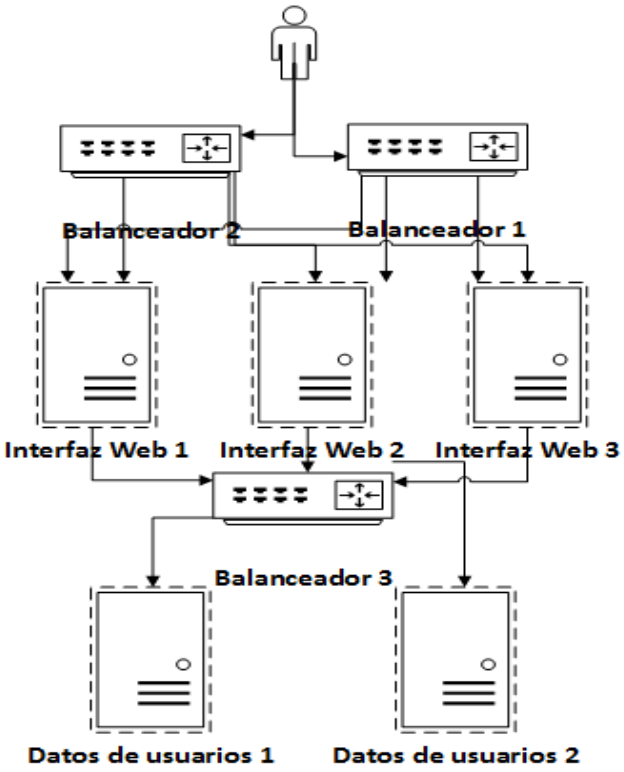

Ilustración 9:Componente de Visualización

La propuesta de hardware para cada servidor es la siguiente:

Cada servidor web: mínimo 4 GB de RAM.

Cada servidor web: mínimo 50 GB de almacenamiento.

Cada servidor web: mínimo un procesador de 2 núcleos.

Para el balanceador de carga (Solr-Web): mínimo 50 GB de almacenamiento.

Para el balanceador de carga (Solr-Web): mínimo 4 GB de RAM.

Para el balanceador de carga (Solr-Web): mínimo un procesador de 2 núcleos.

Para el balanceador de carga (Web-Datos de Usuarios-): mínimo 50 GB de almacenamiento.

Para el balanceador de carga (Web- Datos de Usuarios): mínimo 4 GB de RAM.

Para el balanceador de carga (Web- Datos de Usuarios-): mínimo un procesador de 2 núcleos.

Para el balanceador de carga 1(Usuarios-Con la Web): mínimo 50 GB de almacenamiento.

Para el balanceador de carga 1(Usuarios-Con la Web): mínimo 4 GB de RAM.

Para el balanceador de carga 1(Usuarios-Con la Web): mínimo un procesador de 2 núcleos.

Los servidores de datos de usuarios: mínimo con 4 GB de RAM.

Los servidores de datos de usuarios: mínimo con 300 GB de almacenamiento.

Los servidores de datos de usuarios: mínimo con un procesador de 4 núcleos.

Componente estadístico (Ilustración 10): Este componente se encarga de procesar los datos almacenados en los servidores de Datos de usuarios y los servidores de indexación, del sistema para arrojar estadísticas sobre los contenidos de la web y las búsquedas de los usuarios. La arquitectura que se propone define un servidor para todo el procesamiento estadístico de estos datos. En caso de que la información esté replicada en los servidores de indexación, el balanceador de carga distribuirá el procesamiento al servidor con menos carga de procesamiento.

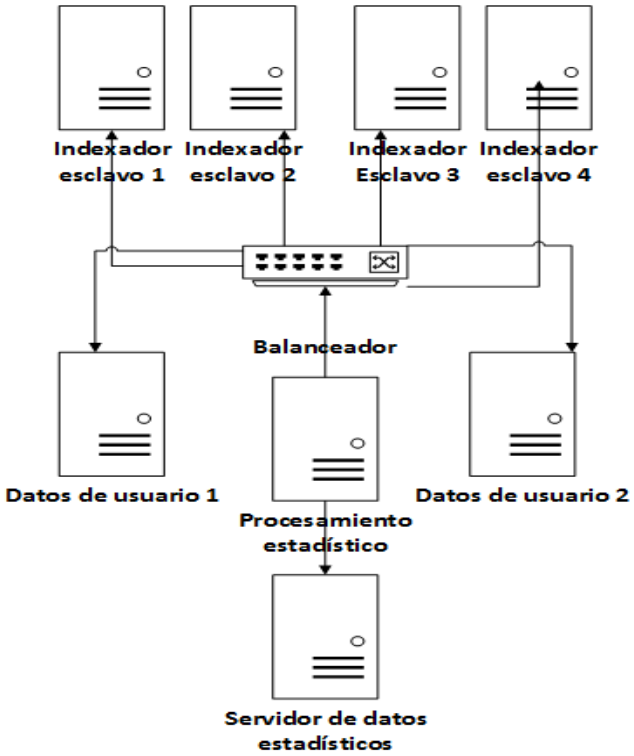

Ilustración 10:Componente Estadístico

La propuesta de hardware para el servidor es la siguiente:

El servidor de datos estadísticos: mínimo 8 GB de RAM.

El servidor de datos estadísticos: mínimo 100 GB de almacenamiento.

El servidor de datos estadísticos: mínimo un procesador de 4 núcleos.

Componente de apoyo a la toma de decisiones (Ilustración 11): Este componente se encarga de brindar mecanismos de reportes que apoyan la toma de decisiones relativas a los contenidos de la web y las búsquedas de los usuarios. La información procesada se convierte en una rica fuente de conocimiento para detectar tendencias de búsqueda de los usuarios de la red y proporcionar así contenido relacionado con sus gustos; además de facilitar datos sobre los perfiles de publicación de cada uno de los sitios web. El resultado final de este componente es una serie de informes estadísticos que manejan datos tales como (Maldonado, y otros, 2017):

Perfil de búsqueda de usuario

Perfil de publicación de sitio

Consultas más buscadas

Buscar preferencias por regiones geográficas o áreas institucionales

Porcentaje de publicación de las categorías de un dominio específico

La arquitectura en la que se basa este componente, está compuesta por un servidor que maneja todo el procesamiento para generar los reportes de apoyo a la toma de decisiones, que extrae los datos a procesar del servidor de datos estadísticos, a los reportes generados se accede a través de los servidores de interfaces web.

La propuesta de hardware para el servidor es la siguiente:

El servidor de apoyo a la toma de decisiones: como mínimo 8 GB de RAM.

El servidor de apoyo a la toma de decisiones: mínimo 300 GB de almacenamiento.

El servidor de apoyo a la toma de decisiones: mínimo un procesador de 4 núcleos.

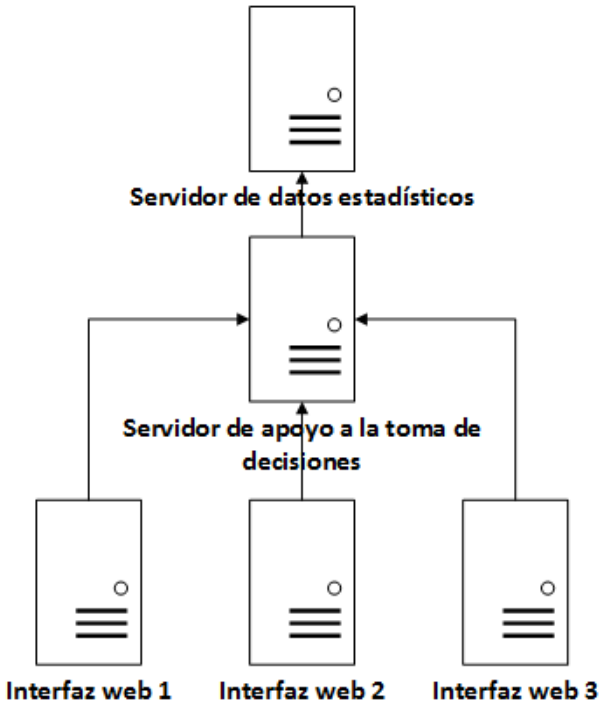

Ilustración 11:Componente de Apoyo a la Toma de Decisiones

2.5.2 Configuraciones de Seguridad Básicas

Las configuraciones de seguridad dependen en gran medida de lo que el administrador o los desarrolladores del buscador deseen asegurar y las políticas que definan para ello. Además, se tiene en cuenta la salida y entrada de información, los puertos, Direcciones IP y protocolos. Para la salida de los servidores nutch a internet se define una dirección IP, mediante el nateo (NAT), la misma replica a otros IP reales pertenecientes a los servidores, es decir que forman parte de la red interna de la arquitectura. Para cada servidor debe definirse una dirección IP, el protocolo y los puertos de red que utilizan. A continuación, se presentan las políticas básicas a tener en cuenta para el despliegue de la arquitectura:

Para los servidores nutch: Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico saliente.

Para los servidores de balanceo de carga (Nutch-Solr): Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante de los servidores de nutch.

Para los servidores solr maestro: Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante del servidor de balanceo de carga (Nutch-Solr).

Para los Servidores solr esclavos: Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante de los solr maestros, los balanceadores de carga (Solr-Web).

Para el servidor de procesamiento: Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante de los servidores de datos de usuarios y de los servidores solr esclavos.

Para los servidores de balanceo de carga (Solr-Web): Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante de los servidores web.

Para el servidor datos estadísticos: Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante del servidor de procesamiento.

Para el servidor de apoyo a la toma de decisiones: Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante del servidor de datos estadísticos.

Para los servidores web: Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante de los servidores de balanceo de carga (Usuarios-Web).

Para los servidores de datos de usuarios: implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante de los balanceadores de carga(DatosUsuarios-Web)

Para los servidores de balanceo de carga (Usuarios-Web): Implementación del firewall con las siguientes políticas:

- Permitir solo el tráfico entrante de consultas de usuarios por el protocolo HTTPS y el puerto 8080.

2.6 Descripción de los estilos arquitectónicos y los patrones de diseño

Como base para el correcto diseño de la arquitectura propuesta, se define el uso de un estilo arquitectónico y patrones de diseño necesarios.

Estilos Arquitectónicos

Hay un buen número de definiciones de componentes, pero Clemens Alden Szyperski proporciona una bastante operativa: un componente de software, dice, es una unidad de composición con interfaces especificadas contractualmente y dependencias del contexto explícitas (Szyperski, 2002).

Como estilo arquitectónico se define utilizar la Arquitectura por componentes, esta se enfoca en la descomposición del diseño en componentes funcionales o lógicos que expongan interfaces de comunicación bien definidas.

Principales beneficios del estilo de arquitectura basado en componentes:

Facilidad de Instalación: Cuando una nueva versión esté disponible, usted podrá reemplazar la versión existente sin impacto en otros componentes o el sistema como un todo.

Costos reducidos: El uso de componentes de terceros permite distribuir el costo del desarrollo y del mantenimiento.

Facilidad de desarrollo: Los componentes implementan una interface bien definida para proveer la funcionalidad definida permitiendo el desarrollo sin impactar otras partes del sistema.

Reusable: El uso de componentes reutilizables significa que ellos pueden ser usados para distribuir el desarrollo y el mantenimiento entre múltiples aplicaciones y sistemas.

Mitigación de complejidad técnica: Los componentes mitigan la complejidad por medio del uso de contenedores y sus servicios. Ejemplos de servicios de componentes incluyen activación, gestión de vida, gestión de colas de mensajes para métodos del componente y transacciones.

Patrones de Diseño

Los patrones de diseño constituyen una serie de estrategias a seguir, que facilitan la descripción del problema y su solución. “Los patrones de diseño son el esqueleto de las soluciones a problemas comunes en el desarrollo de software” (Tedeschi, 2014).

Los patrones son utilizados para lograr un software flexible, capaz de tener componentes reutilizables. Es un lenguaje entre los diseñadores, capaz de trasmitir a través de experiencias el mejor patrón a seguir ante un determinado problema.

Dentro de los patrones arquitectónicos se encuentran:

- Modelo Vista Controlador(MVC): es uno de los modelos más antiguos (*Smalltalk*-80) y por lo tanto se convirtió en uno de los patrones fundamentales para el desarrollo de software. MVC a grandes trazos, separa las preocupaciones con respecto a los datos (modelo) y la interfaz de usuario (vista/GUI), permitiendo modificaciones independientes en cada una las partes sin afectar la otra, o sea, para que los cambios realizados en la interfaz de usuario no afectan el manejo de datos, y los datos pueden ser reorganizados sin cambiar la interfaz de usuario.
- Inyección de Dependencias: es un patrón que a pesar de ser relativamente nuevo es muy complejo. La utilización de inyección de dependencia en un proyecto es tan incidente, que puede modificar en grandes proporciones la arquitectura, de modo que se hace prudente una planificación a futuro sobre la utilización de este patrón. Este patrón puede dar la impresión de ir un poco “al revés”, se trata de una aplicación de la “Inversión de Control – IoC”, concepto que hace exactamente eso, se invierte el flujo de control. El nombre de “la inyección de dependencia” en realidad se presta a confusión, ya que el modelo permite que se inyecte no dependencias en sí, sino en su lugar, la información para satisfacerlas la relación a las dependencias.
- Arquitectura dirigida por eventos (*Event-driven architecture o EDA*): es un patrón de arquitectura software que para orquestar su comportamiento se centra en torno a la producción, detección, consumo y respuestas ante “eventos”. Teniendo en cuenta que un evento es: cualquier ocurrencia identificable que tiene un significado para el hardware o el software del sistema. Y a su vez este cambio de estado puede ser conocido por otras aplicaciones en la arquitectura, o sea, que cada evento se propaga de manera inmediata a otras partes del sistema en la medida que sea necesario.
- Arquitectura orientada a servicios(SOA): La ‘Arquitectura Orientada a Servicios de cliente (*Service Oriented Architecture)*, es un concepto de arquitectura de software donde el software consta de una composición de servicios, prestaciones y reglas, y son los requisitos del negocio los que dictaminan la manera en la que estas se ínter-relaciona. Está diseñado para que el sistema sea altamente escalable y flexible a nuevos requerimientos.

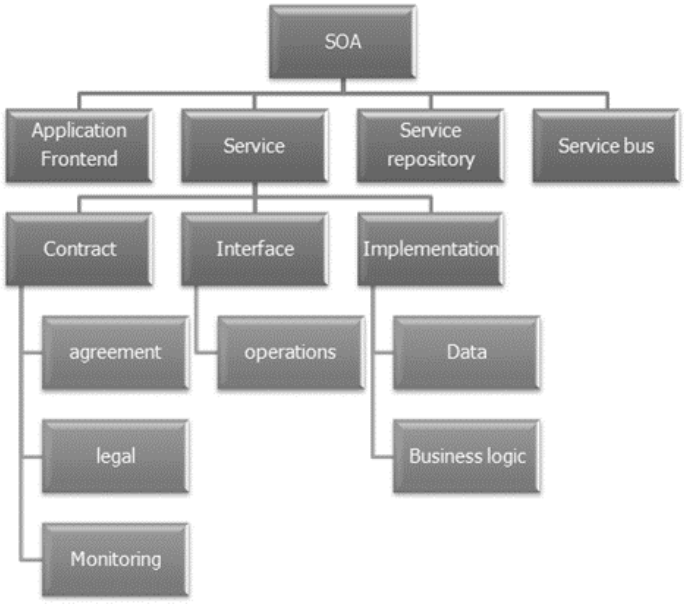

Ilustración 12:Diseño de Arquitectura SOA(López, 2007)

Como arquitectura se propone la Arquitectura orientada a servicios debido a que posee las siguientes características:

Aumentar la eficiencia, agilidad y productividad de un grupo de trabajo, haciendo hincapié en los servicios como el principal medio.

Exponer procesos de negocio como servicios es la clave a la flexibilidad de la arquitectura. Esto permite que otras piezas de funcionalidad (incluso también implementadas como servicios) hagan uso de otros servicios de manera natural, sin importar su ubicación física.

La Arquitectura orientada a servicios resultante, define los servicios con los que estará compuesto el sistema, sus interacciones, y con qué tecnologías serán implementados (Tobar, 2015).

Conclusiones del Capítulo

Luego de abordar de forma breve la fundamentación acerca de la propuesta de solución descrita por el autor, así como, haber definido todos los parámetros necesarios para lograr un correcto análisis y diseño de la solución, podemos concluir que:

- A partir del estudio realizado en el presente capítulo, se definieron como componentes necesarios para lograr el correcto desarrollo de la solución:
  Componente de rastreo
  Componente de indexación
  Componente de procesamiento
  Componente de estadística
  Componente de apoyo a la toma de decisiones
  Componente de visualización
- Con la especificación de los requerimientos, se lograron definir todos los parámetros con los que debe cumplir el diseño, seleccionándose como los requerimientos más importantes la: gestión de hardware de los componentes y proteger los datos basándose en la confidencialidad, integridad y disponibilidad.
- La definición de un estilo arquitectónico, y patrones de diseño descritos, permitió estructurar una arquitectura acorde a las necesidades del sistema y características de la propuesta de solución.
- La propuesta cuenta con 25 servidores entre los que se encuentran: 4 servidores de rastreo, 6 servidores de indexación, 2 maestros y 4 esclavos, 3 servidores de visualización, 1 servidor de procesamiento, 1 servidor de datos estadísticos, 2 servidores de datos de usuarios, 1 servidor de apoyo a la toma de decisiones y 7 servidores de balanceo de carga distribuidos por toda la arquitectura.

CAPÍTULO 3: VALIDACIÓN PARA LA PROPUESTA DE ARQUITECTURA Y CONFIGURACIONES DE SEGURIDAD PARA EL BUSCADOR ORIÓN V2

3.1 Introducción

En el siguiente capítulo se definen las pruebas a realizar, se describen los procesos asociados al despliegue del sistema y se da un seguimiento a través de pruebas verificando validez a los requerimientos y garantizar el óptimo funcionamiento de la aplicación hacienda una valoración crítica a partir de los resultados obtenidos.

3.2 Pruebas realizadas para la validación de la arquitectura propuesta

Para la validación de la hipótesis científica que guía el proceso de investigación, se hace necesario el uso de pruebas que validen la propuesta de solución que se expone, teniendo en cuenta el rendimiento y la seguridad de dicha arquitectura.

3.3.1 Prueba de Rendimiento

Con la realización de las pruebas de rendimiento se analiza el comportamiento de los servidores y las deficiencias que generan el bajo rendimiento. Este tipo de pruebas permiten identificar cuellos de botella, capacidad de concurrencia de usuarios, tiempos de respuesta de operaciones de negocio a nivel de sistema, establecer un marco de referencia para pruebas futuras, determinar el cumplimiento de los objetivos de rendimiento, entre otros (V&v Quality S.A., 2015).

En esta prueba se ejecutan las pruebas de carga y estrés que analizamos a continuación:

- Pruebas de estrés: Permite identificar la capacidad de respuesta de un sistema bajo condiciones de carga extrema, representadas por una alta concurrencia de Usuarios y/o procesos. Lo que posibilita asegurar una mejor capacidad de concurrencia de usuarios y/o procesos que se verá reflejada en una óptima operación de negocio (V&v Quality S.A., 2015).
- Pruebas de carga: Permite identificar la capacidad de recuperación del sistema cuando es sometido a cargas variables tanto de usuarios como de procesos. Con la ejecución de estas pruebas se puede determinar el tiempo de respuesta de todas las transacciones críticas del sistema y encontrar cuellos de botella (V&v Quality S.A., 2015).

Resultado de las pruebas de rendimiento

Para las pruebas de rendimiento se utiliza la herramienta Grafana 4.1, encargada del monitoreo de los servidores. Se decide como entorno de trabajo la Plataforma de Servicios C.U.B.A, donde se analizan los componentes fundamentales del buscador: rastreo, indexación y visualización. Se recogen y comparan los resultados obtenidos en tiempo real.

- En el caso de los **servidores de rastreo**: Nutch

Cada servidor nutch como promedio, rastrea diario alrededor de 20 000 páginas, según los administradores del sitio. Además, cada uno cuenta con varios *spiders* que se ejecutan por 3 rondas y cada ronda posee 8 etapas para ejecutar un rastreo, enumeradas a continuación: *1-Injector, 2-Generator, 3-Fetching, 4-Parcing, 5-CrawlDb, 6-LinkDb, 7-Indexer, 8-Cleanup.* Como punto de partida poseen un semillero de alrededor de 30 URL, para comenzar el análisis.

Para demostrar el correcto funcionamiento de los servidores al aplicar la propuesta de solución, se realiza una comparación del estado del rendimiento antes y después de aplicar las configuraciones propuestas.

Antes de aplicar la propuesta

Se presenta el hardware de los servidores teniendo en cuenta: CPU, RAM y almacenamiento de disco duro, actual del buscador.

Hardware de los servidores nutch de la Plataforma de Servicios C.U.B.A:

CPU: 4 núcleos

RAM: 3 GB

Almacenamiento: 300 Gb

Ejecución de las pruebas con la arquitectura propuesta

A continuación, se presentan los 4 servidores con sus respectivos spiders y la arquitectura rediseñada:

Tabla 3: Cantidad de spiders por servidores nutch

Servidores Nutch	*Spiders* por Servidor
Cuba13	• Prensa • Imágenes-Prensa

	• Dedicados
Cuba 14	• Infomet(Sitio de Salud Cubano) • Imágenes-Infomet • Imágenes-Dedicadas
Cuba 15	• Cultura • Imágenes-Cultura
Cuba 16	• Educación • Imágenes-Educación • General • Imágenes-General • Thumbnails-Dedicado-Cubadebate

Para los servidores Cuba13, Cuba14, Cuba15 se presenta:

CPU: 4 núcleos

RAM: 4 Gb

Almacenamiento: 300 Gb

En el caso del servidor Cuba16 se decide aumentar la RAM con el objetivo de analizar su comportamiento con un mayor número de *spiders*, previendo el aumento de los mismos con el tiempo:

CPU: 4 núcleos

RAM: 6 Gb

Almacenamiento: 300 Gb

A continuación, se muestra una tabla donde se observa cómo se comportan los servidores antes y después de la aplicación de las configuraciones propuestas en la solución:

Dado a la cantidad de prestaciones que varían entre los servidores y con el objetivo de comparar las diferencias entre ellos, se presenta una tabla donde se analiza el rendimiento antes y después de aplicados los cambios:

Tabla 4: Evaluación de recursos antes y después en los servidores de rastreo

Servidor Nutch: Cuba13	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 83%	CPU: 1%
RAM: 98%	RAM: 26%
DD: 65%	DD: 35%
Servidor Nutch: Cuba14	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 75%	CPU: 10%
RAM: 98%	RAM: 25%
DD: 65%	DD: 32%
Servidor Nutch: Cuba15	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 58%	CPU: 2%
RAM: 98%	RAM: 21%
DD: 30%	DD: 28%
Servidor Nutch: Cuba16	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 86%	CPU:88%
RAM: 99%	RAM: 54%
DD: 65%	DD: 34%

En la tabla se observa como sin las configuraciones propuestas se tiene un elevado uso del CPU y RAM lo que demuestra que, con las configuraciones de hardware que presentaba no se obtenía un buen rendimiento, además de que los servidores se encontraban expuestos a sobrecargas, lo que puede provocar fallos en su funcionamiento. Luego de la aplicada la propuesta se perciben cambios

positivos en general, demostrando que con una mejor distribución de los recursos destinados a los *spiders* se pueden lograr grandes mejoras en el buscador. Por lo que se puede concluir que la arquitectura propuesta ayudará en la eficiencia y el desempeño del buscador. Esto se debe a un mejor funcionamiento de los servidores, los balanceadores y la carga de trabajo entre ellos.

- En el caso de los **servidores de indexación:** Solr

Estos servidores se encargan de la indexación de la información y el manejo de los datos, ya sea en el aumento, actualización o eliminación de documentos.

Antes de aplicar la propuesta

A continuación, se analiza el uso de los recursos de un servidor maestro y un esclavo sin aplicar la propuesta de solución:

Hardware de un servidor solr maestro:

CPU: 4 núcleos

RAM: 8 Gb

Almacenamiento: 100 Gb

Hardware de un servidor solr esclavo:

CPU: 2 núcleos

RAM: 5 Gb

Almacenamiento: 100 Gb

Al poseer tales recursos, no se presentaba un gran uso de los servidores, y esto se debe al poco rastreo de información que proveniente de los servidores nutch, lo que generaba poco volumen de páginas a indexar. Esto provocaba baja actualización de la información, además de una baja aceptación de los usuarios hacia el buscador.

En un mundo donde la web está en constante evolución, cada día son más las páginas, sitios, dominios y documentos a indexar, por lo que se hace necesario la propuesta reforzar las prestaciones existentes.

Ejecución de las pruebas con la arquitectura propuesta

Para la ejecución de la prueba se presenta la nueva propuesta de hardware a aplicar:

Hardware de un servidor solr maestro:

CPU: 2 núcleos

RAM: 8 Gb

Almacenamiento: 200 Gb

Hardware de un servidor solr esclavo:

CPU: 4 núcleos

RAM: 4 Gb

Almacenamiento: 100 Gb

Solr indexa actualmente alrededor de 50 000 o 55 000 páginas diarias ya sean actualizaciones, nuevos documentos o eliminaciones, con un promedio de incremento de 7000 páginas diarias.

Para una mejor comprensión de la distribución de recursos de los servidores, se presenta una tabla con la información, perteneciente al antes y después de los cambios en los servidores:

Tabla 5: Evaluación de recursos antes y después del rendimiento de los servidores de indexación

Servidores Solr Maestro	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 0.8%.	CPU: 63%
RAM: 25%	RAM: 60%
DD: 24%	DD: 24%
Servidores Solr Esclavo	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 1.5%	CPU: 22%
RAM: 53%	RAM: 53%
DD: 33%	DD: 35%

Una vez aplicada las configuraciones propuestas al componente de indexación se evidencia un mayor uso de los recursos del servidor, en comparación con lo presentado en anteriormente. Esto se debe al aumento de documentos recibidos para indexar, debido a las mejoras realizadas al componente de rastreo. Al mejorar el funcionamiento de los servidores nutch, estos envían un mayor número de información hacia los servidores solr, lo que conlleva al procesamiento de más información y más rápido, de ahí el consumo de recursos evidenciado. Esto garantiza el incremento de la información

almacenada y su actualización. Ante lo expuesto se puede concluir que el aumento en el rendimiento de estos servidores influye de forma positiva para el buscador.

- En el caso de los **servidores de visualización**: Nginx

Esta prueba se desarrolla en las Interfaces de los servidores Web, teniendo en cuenta el tiempo de respuesta de una consulta realizada y la cantidad de usuarios concurrentes en el sitio. Mediante un análisis de datos realizado tanto al sitio, como a su interfaz, se arriba a la conclusión, que el mismo cuenta con un tiempo de respuesta de 158ms aproximadamente en una consulta y la interfaz responde en un promedio de 0.203 segundos.

Antes de aplicar la propuesta

Para una mejor comprensión del análisis de la prueba desarrollada en las Interfaces de los servidores Web, se presentan los recursos de hardware que posee la arquitectura y el rendimiento de la misma sin aplicar las configuraciones propuestas, con 20 usuarios concurrentes. Posteriormente se analizarán con algunos cambios propuestos y así comparar las mejoras efectuadas a la arquitectura.

Hardware de los Servidores Web:

CPU: 2 Núcleos

RAM: 2 Gb

Almacenamiento: 50 Gb

Hardware de los Balanceadores de Carga de los Servidores Web:

CPU: 2 Núcleos

RAM: 3 Gb

Almacenamiento: 10 Gb

Ejecución de las pruebas con la arquitectura propuesta

El resultado se evidencia con el monitoreo de los recursos de los servidores de balanceo de carga y los webs, con 100 usuarios concurrentes:

Con el fin de brindar respuestas más rápidas a los usuarios se propone para los Servidores Web y los Servidores de Balanceo de Carga de la Plataforma de Servicios C.U.B.A:

CPU: 2 Núcleos

RAM: 4 Gb

Almacenamiento: 50 Gb

A continuación, se presenta una tabla, donde se evidencia el rendimiento antes y después de aplicar la propuesta de solución.

Tabla 6: Evaluación de recursos antes y después del rendimiento de los servidores de aplicación

Servidores Web 3	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 0.9%	CPU: 1.3%
RAM: 21%	RAM: 23%
DD: 21%	DD: 21%
Servidor Web 4	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 1.0 %	CPU: 1.3%
RAM: 20%	RAM: 24%
DD: 29%	DD: 29%
Servidor Web 5	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 1.0%	CPU: 1.2%
RAM: 23%	RAM: 25%
DD: 24%	DD: 24%
Balanceador de Carga 1	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 2%	CPU: 1.0%
RAM: 23%	RAM: 20%
DD: 24%	DD: 22%

Balanceador de Carga 2	
Sin aplicar las configuraciones propuestas	Aplicadas las configuraciones
CPU: 2%	CPU: 1.0%
RAM: 24%	RAM: 22%
DD: 26%	DD: 22%

Una vez analizadas la tabla expuesta acerca del monitoreo del componente de visualización, se concluye que con una diferencian de 80 usuarios y algunos cambios realizados en la distribución de los recursos de los servidores, se obtiene un buen rendimiento de estos. Al mejorar los recursos, se gana en la aceptación de los usuarios, y la velocidad del buscador, ya que genera consultas y respuestas más rápidas. También proporciona mayor número de almacenamiento de los registros o *logs* de la aplicación y de los servicios que esta brinda.

3.3.2 Pruebas de Seguridad

Las pruebas de seguridad garantizan que los usuarios estén restringidos a funciones específicas o que su acceso esté limitado. Sólo aquellos usuarios autorizados por el sistema son capaces de ejecutar las funcionalidades disponibles. El objetivo fundamental de este tipo de pruebas es comprobar los niveles de seguridad lógica del sistema. Para la realización de esta prueba se utiliza la herramienta Acunetix Web Vulnerability Scanner 8.0.

Una vez ejecutadas las pruebas de seguridad es posible detectar las vulnerabilidades a las que se ve expuesto el sistema, así como tomar medidas para la disminución de amenazas de ataques (Torets, 2016).

Resultados de las pruebas de seguridad

Tabla 7: Resultado de las pruebas de seguridad

Categorías de vulnerabilidades	Cantidad de Errores
Denegación de servicios	0
Transición insegura de HTTP a HTTPS	0
Integridad de la información	0

Total	0

Transición insegura de HTTP a HTTPS: Esta vulnerabilidad permite a un atacante reemplazar el destino del formulario.

Denegación de servicios: Esto provoca que los usuarios no accedan a la información deseada.

Integridad de la información: Esta forma de ataque conlleva a que la información almacenada se pueda ver afectada, ya sea modificada o eliminada.

Una vez realizadas las pruebas de rendimiento y seguridad, teniendo en cuenta la asignación, distribución y el análisis de las vulnerabilidades ante ataques definidas, se puede concluir, que un correcto diseño de arquitectura de hardware y configuraciones de seguridad, permitirá grandes mejoras en el rendimiento y la seguridad del buscador Orión v2.

3.4 Validación de la Hipótesis de la Investigación

La validación basada en el juicio de expertos permite obtener valoraciones sobre temas relacionados con la propuesta de solución. Como método a utilizar en el procesamiento estadístico de estos criterios, fue aplicada la escala psicométrica creada por Rensis Likert en 1932 (Boone, 2012). La misma fue utilizada en esta investigación a través de un cuestionario para conocer el nivel de acuerdo y desacuerdo con los objetivos y componentes de la arquitectura de hardware y configuraciones de seguridad propuestas.

Criterio de Experto-Escalamiento de Likert

Para el desarrollo de esta validación se definen además las personas que, a criterio del autor cumplen los requisitos de especialistas o trabajadores que se encuentran relacionados con la base teórica y práctica de la investigación. Se seleccionan 10 expertos pertenecientes a la Facultad 1, que se encuentran vinculados a áreas relevantes para el desarrollo de la investigación. Se tuvieron en cuenta los siguientes aspectos: grado científico, categoría docente, años de experiencia en el área, nivel de dominio sobre el tema que se encuesta. Todos cumplen los requisitos de expertos y tienen experiencia en actividades vinculadas al buscador Orión. La caracterización de los expertos es: 10 ingenieros, de ellos 3 master, todos con más de 2 años de experiencia en el área y vinculados a la producción y la docencia.

Una vez analizados los especialistas seleccionados, se presentan los indicadores definidos para las valoraciones de la arquitectura (Tabla 8), además de los aspectos a tener en cuenta en el cuestionario y

posteriormente se procesan los resultados mediante la Escala de Likert (Tabla 9) o a través de la (Ilustración 13).

Los expertos expresan sus valoraciones mediante indicadores como se presentan en la siguiente tabla:

Tabla 8: Indicadores para la validación de la hipótesis científica

5	4	3	2	1
Totalmente de acuerdo	De acuerdo	Ni de acuerdo ni en desacuerdo	En desacuerdo	Totalmente desacuerdo

Con esta técnica son calculados los porcientos de concordancia de los expertos en las respuestas dadas a los planteamientos definidos en la encuesta. Luego se calcula el índice porcentual (IP) que integra en un valor la aceptación de cada planteamiento evaluado mediante la siguiente fórmula:

IP = 5(%)+4(%)+3(%)+2(%)+1(%) / 5

Los principales aspectos utilizados para aplicar Likert fueron los siguientes:

- La estructura propuesta para la arquitectura cubre los componentes fundamentales de un SRI.
- La arquitectura es escalable.
- La arquitectura maneja de forma correcta la variable disponibilidad de los datos.
- La arquitectura maneja de forma correcta la variable balanceo de carga.
- La arquitectura maneja de forma correcta la variable respaldo de los datos.
- La arquitectura propone configuraciones de seguridad básicas para sistema.
- Las configuraciones de seguridad propuestas protegen la integridad de los datos.

A continuación, se presenta el resultado de la validación por criterio de expertos utilizando el escalamiento de Likert. El mismo propone mediante el cálculo en porciento los valores de las preguntas analizadas, teniendo en cuenta los indicadores propuestos (Tabla 8).

Para la obtención de valores se analizó las respuestas de los expertos teniendo en cuenta que: 3 de los expertos estuvieron totalmente de acuerdo con todas las afirmaciones, 4 de ellos estuvieron de acuerdo

en una de las afirmaciones y totalmente de acuerdo con el resto y los 3 expertos restantes estuvieron de acuerdo en dos afirmaciones y totalmente de acuerdo en el resto de las afirmaciones.

Tabla 9: Valoración de expertos sobre la propuesta

ESCALA						
Preguntas	**IP**	**TD**	**DA**	**NI**	**ED**	**TA**
P1	100	100	0	0	0	0
P2	100	100	0	0	0	0
P3	92	60	40	0	0	0
P4	90	50	50	0	0	0
P5	100	100	0	0	0	0
P6	100	100	0	0	0	0
P7	98	90	10	0	0	0

Para un mejor entendimiento de la tabla analizada anteriormente acerca de la valoración de experto, se presenta el siguiente gráfico de barras:

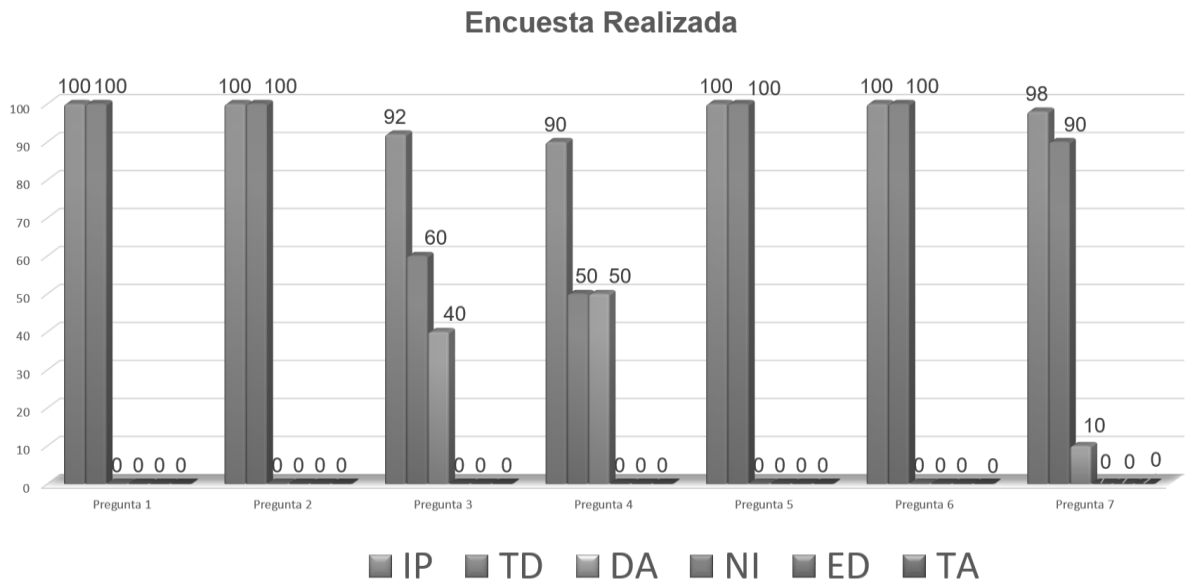

Ilustración 13: Valoración de Expertos-Criterio de Likert

Con la realización del escalamiento de Likert se evidencia que tanto los elementos teóricos, como los prácticos, tienen una buena valoración por parte de los especialistas seleccionados. Durante el proceso

se constataron criterios favorables para el diseño de la arquitectura propuesto. Se concluye que el índice porcentual por preguntas supera en cada caso el 90 %, lo que indica que hay un acuerdo entre los expertos con relación a la calidad de la propuesta.

Conclusiones del Capítulo

Una vez desarrollado el presente capítulo acerca de la validación de la propuesta de diseño de arquitectura de hardware y sus configuraciones de seguridad, teniendo en cuenta el resultado de las pruebas definidas en el mismo, se presentan las siguientes conclusiones:

- La monitorización de los componentes de hardware es un aspecto fundamental en la administración de SRI.
- La aplicación de las pruebas de rendimiento y seguridad, permitieron identificar las principales deficiencias de la arquitectura actual y a su vez generar una estrategia para solucionar las mismas, obteniendo una propuesta con mayor calidad y rendimiento con el menor gasto de recursos posible.
- Mediante la técnica del Criterio de Expertos-Escalamiento de Likert se constató la validez de la propuesta de solución a través de un cuestionario, teniendo gran relevancia para la toma de decisiones en cuanto al mejoramiento de la arquitectura.

CONCLUSIONES

Una vez finalizada la fundamentación teórica que sustentó la presente investigación, definidas las características del sistema, así como efectuado su desarrollo y validación, se obtuvieron resultados que permiten arribar a las siguientes conclusiones:

- El estudio de los SRI analizados permitió generalizar que los principales componentes para el funcionamiento de un buscador son: rastreo, indexación y visualización. Además, se propuso incorporar a Orión v2 componentes para el procesamiento, el análisis estadístico y el apoyo a la toma de decisiones.
- La investigación realizada permitió identificar los principales requerimientos que contribuyeron al diseño de la arquitectura y configuraciones de seguridad del buscador Orión v2.
- La arquitectura propuesta para sistemas de recuperación de información facilita el proceso de réplica de datos, lo que propicia así, establecer mecanismos de respaldo ante fallas.
- La organización distribuida de los servidores, mejora el tiempo de respuesta a los usuarios, optimiza el almacenamiento y procesamiento de los datos de forma general; además permite aumentar el número de servidores para crecer en capacidad de rastreo y almacenamiento o disminuir la cantidad de servidores para ahorrar en costos de hardware sin afectar el funcionamiento general del sistema.
- Con el diseño de la arquitectura y configuraciones de seguridad del buscador Orión v2, se obtuvo una propuesta aplicable a cualquier entorno web, por lo que puede ser utilizada para el desarrollo y despliegue de sistemas de recuperación de información a nivel nacional o internacional.
- La aplicación de las pruebas realizadas permitió constatar que la propuesta realizada permite mejorar el rendimiento y la seguridad del motor de búsqueda Orión v2.

RECOMENDACIONES

- Incorporar a la arquitectura que se propone los *backup* para la prevención de cortes de energía.
- Agregar a la arquitectura un Servidor de registros, para un mejor funcionamiento de los servidores Web.

BIBLIOGRAFÍA

Definición ABC. 2015. [En línea] 4 de 10 de 2015. http://www.definicionabc.com/tecnologia/buscador.php.

Visual Paradigm. 2014. .Visual Paradigm for UML. *Visual Paradigm for UML - Software design tools for agile software development.* [En línea] 2014. http://www.visualparadigm.com/product/vpuml/.

Abadal, E. y CODINA, L. 2005. Recuperación de Información. Bases de Datos Documentales: Características, funciones y método. 2005.

Antiñanco, M. J. 2014. *Bases de Datos NoSQL: escalabilidad y alta disponibilidad a través de patrones de diseño.* s.l. : Tesis Doctoral. Facultad de Informática., 2014.

Apache Software Fundation. 2015. : The Apache Solr Reference Guide. [En línea]. Apache Solr – Resources. [En línea] 2015. http://lucene.apache.org/solr/documentation.html.

Baeza-Yates, R. Ribeiro-Neto, B. 1999. Modern InformationRetrieva l. New York : Essex: Addison-Wesley Longman, 1999.

Balkhi, S. 2013. What is: Apache. 2013.

Benavides, M.Sc. K.D.R. 2015. Arquitectura de Google. 2015.

Bertolín, Jr A. 2008. Seguridad de la información. Redes, informática y sistemas de información. s.l. : Paraninfo, 2008.

Boone, H. N. y Boone, D. A. 2012. *Analyzing likert data.* s.l. : Journal of extension , 50(2), 1 -5., 2012.

CIDI, Centro de Ideoinformática. 2017. Arquitectura de Los Buscadores. La Habana : s.n., 2017.

Cleverdon, C. 1997. The Cranfield tests on index language devices. Links. 1997. Vols. vol. 942, p. 42.

Consoft. 2002. Consoft, ¿Qué son los metabuscadores? [En línea] 2002. http://www.consoft.es/noticias/news_text.asp?id=33219.

Cubanic. 2017. CUBANIC. [En línea] 2017. http://cubanic.cu.

Fonollosa, A. B., L.Díaz y J.Huerta. 2012. Construyendo un sistema de indexación y búsqueda de recursos georreferenciados. 2012.

Grafana. 2017. Grafana. [En línea] 2017.

Guiluz, J. 2013. Desarrollo web ágil con Symfony2. 2013. Vol. pág. 618.

Gutiérrez, M.Pérez-Montoro y Codina, L. 2010. Software de prototipado para la arquitectura de la información: funcionalidad y evaluación. El Profesional de la Informacion. 2010. Vol. vol. 19, , num. 4.

Internet live Stats. 2017. Internet Live Stats. [En línea] 2017. http://www.internetlivestats.com.

Jaimes, L. G. y Vega Riveros, F. 2005. Modelos clásicos de recuperación de la información. 2005. Vol. 23, 1.

Kholodkov, V. 2015. Nginx Essentials. Packt Publishing Ltd. 2015.

Kowalski, G. 1998. Information retrieval systems: theory and implementation, Computers and Mathematics with Applications. 1998. Vol. vol. 5, No. 35, pp. 133,.

Leyva, P. R.z, SALA, H. V. y FLORES, L. A. P. 2016. Componentes y funcionalidades de un sistema de recuperación de la información. s.l. : Revista Cubana de Ciencias Informáticas, 2016. Vols. vol. 10, p. 150-162.

López, Gustavo. 2007. Una propuesta de modelos de ciclo de vida (mcvs) para la integración de los procesos de negocio utilizando Service Oriented Architecture (SOA). En IX Workshop de Investigadores en Ciencias de la Computación. : s.n., 2007.

Lupa. 2017. Lupa. [En línea] 2017. (lupa.upr.edu.cu).

Maldonado, C.O., y otros. 2017. Computational model for the processing of documents and support to. s.l. : nternational Research Journal of Engineering and Technology (IRJET), 2017. Vol. 4.

Molina, B. C. 2015. *Integración de un proxy inverso NGINX con un panel de control Virtualmin para crear una plataforma de servicios de hospedaje Web. . Tesis Doctoral.* 2015.

Nedelcu, C. 2010. Nginx HTTP Server. s.l. : Packt Publishing. ISBN 978-1-84951-086-8., 2010.

Netcraft. 2013. NETCRAFT. Servidores web más usados desde1995. *Sitio web de NetCraft.* [En línea] 2013. http://www.netcraft.com.

Nginx. 2015. [En línea] 2015. https://www.nginx.com/resources/wiki/..

NutchWiki. 2015. “NutchTutorial - Nutch Wiki.”. [En línea] 2015. https://wiki.apache.org/nutch/NutchTutorial. PINTO.

Ortega, M. 2014. Acunetix Web Vulnerability Scanner. [En línea] 2014. https://hakin9.org/acunetix-web-vulnerability-scanner/..

Pastor, J. J. C. 2016. Motores de búsqueda y derechos de autor: infracción y responsabilidad. S.l.: Aranzadi. 2016.

Pinto, M. 2015. Búsqueda y Recuperación de Información. [En línea] 13 de diciembre de 2015. http://www.mariapinto.es/e-coms/busqueda-y-recuperacion-de-informacion/..

Pressman, R. S. 2006. *Ingeniería de Software: Un enfoque práctico.* México : s.n., 2006.

Reinoso, Q. y Paulina, V. 2007. . Definición e implementación de un modelo de respaldos de información en la compañía Transelectric SA. 2007.

Salta, E. F. 2015. Integridad de datos en sistemas de gestión de aprendizaje. En XVII Workshop de Investigadores en Ciencias de la Computación (2015). 2015.

Salton, G. y McGill, M. H. 1983. Introduction to Modern Information Retrieval. New York: McGraw-Hill : s.n., 1983.

Sánchez, S. 2015. *Estrategia de soporte técnico para el proceso de migración a código abierto en los Organismos de la Administración Central del Estado.* La Habana : Universidad de las Ciencias Informáticas, 2015.

Saz, J.T. 2003. Clasificaciones y portales temáticos especializados. Estudio en recursos de información digital sobre ciencias sociales. Salamanca : Tendencias de investigación en organización del conocimiento: En: J.A. Frías y C. Travieso (edito-res.), 2003.

Seta, L. 2010. pache Solr: una introducción. *Apache Solr: una introducción - Dos Ideas.* [En línea] 2010. http://www.dosideas.com/noticias/java/913-apache-solr-una- introduccion.

Sinisterra, M. M., Díaz, T. M. y Ruiz, E. G. 2012. Clúster de balanceo de carga y alta disponibilidad para servicios web y mail. Informador técnico. 2012. no 76, p. 93-102.

Suárez, L. 2016. Buscadores Web. [En línea] 2016. http://buscadoresweb.com/ranking-buscadores-internet/..

Szyperski, C. A. 2002. Component software: Beyond Object-Oriented. 2002.

Tamayo, A. y M.Martinez. 2011. Interoperabilidad de sistemas de organización del conocimiento: el estado del arte. Información, cultura y sociedad. 2011. no 24, p. 15-37.

Tedeschi, N. 2014. Microssoft Developer Network. *Microsoft.* [En línea] 1 de 5 de 2014. http://msdn.microsoft.com/es-es/library/bb972240.aspx..

Tobar, M. Á. BOTTO. 2015. SOA2Cloud: Un marco de trabajo para la migración de aplicaciones SOA a Cloud siguiendo una aproximación dirigida por modelos. 2015.

Torets, J. M. Lopéz. 2016. V&V Quality. [En línea] 18 de mayo de 2016. http://vyvquality.com/pruebas-seguridad/..

V&v Quality S.A. 2015. [En línea] 2015. http://vyvquality.com/pruebas-rendimiento/..

Vilca, E. A. 2017. Academia.edu. [En línea] 2017. http://www.academia.edu/10686259/ARQUITECTURA_DE_HARDWARE_Y_SOFTWARE.

Yahoo!! 2016. Yahoo!! [En línea] 2016. http://www.yahoo.com.

Zambrano, R. A. 2015. Metodología de la investigación. *Metodología de la investigación.* [En línea] 22 de 10 de 2015. http://es.slideshare.net/Mal3J1Ta/resumen-capitulos1234-del-libro.