UNIVERSIDAD CENTROCCIDENTAL

"LISANDRO ALVARADO"

PROPUESTA DE UNA ARQUITECTURA DE SOFTWARE PARA LA CONFORMACION DE UN REPOSITORIO DE DATOS DE REDES SOCIALES PARA ANÁLISIS Y MONITOREO

MARIA ESPERANZA LINAREZ ARTEAGA

Barquisimeto, Abril 2015

UNIVERSIDAD CENTROCCIDENTAL “LISANDRO ALVARADO”
DECANATO DE CIENCIAS Y TECNOLOGÌA
POSTGRADO EN CIENCIAS DE LA COMPUTACIÓN

PROPUESTA DE UNA ARQUITECTURA DE SOFTWARE PARA LA CONFORMACION DE UN REPOSITORIO DE DATOS DE REDES SOCIALES PARA ANÁLISIS Y MONITOREO

Trabajo presentado para optar al grado de
Magíster Scientiarum

Por: MARIA ESPERANZA LINAREZ ARTEAGA

Barquisimeto, Abril 2015

PROPUESTA DE UNA ARQUITECTURA DE SOFTWARE PARA LA CONFORMACION DE UN REPOSITORIO DE DATOS DE REDES SOCIALES PARA ANÁLISIS Y MONITOREO

Por: MARIA ESPERANZA LINAREZ ARTEAGA

Trabajo de grado aprobado

_____ _____

(Jurado 1) (Jurado 2)

_____

(Jurado 3)

Barquisimeto, Abril 2015

DEDICATORIA

A Dios que me ha dado la vida, salud, sabiduría, fortaleza y siempre ha estado a mi lado para culminar este trabajo.

A la Divina Pastora porque su bendición siempre estuvo presente en la ejecución de este trabajo.

A mis padres Ernesto Linarez y Carmen Arteaga de Linarez quienes me han brindado apoyo incondicional, compensación, paciencia y fuerza en todo este tiempo.

A mi tía Violeta Linares por su constante cooperación, apoyo ilimitado y la constante ayuda que siempre me brinda.

A mi querido hermano Ernesto Linarez, porque siempre me da fuerza y cree en mí para culminar mis metas.

A mi abuela Blanca Violeta de Linares, por su apoyo y oraciones que siempre me acompañan.

A mi abuela Rosa María Noguera, que aunque no esté con nosotros físicamente, sus consejos de perseverancia inculcados en mi niñez siempre son base para el logro de mis metas.

AGRADECIMIENTOS

Agradezco a Dios todopoderoso por estar conmigo en todos los momentos de mi vida, darme fe, esperanza, fortaleza, vida y salud para culminar este objetivo, su bendición me ha permitido obtener amor, gozo, paz, paciencia, benignidad y fe para afrontar los momentos difíciles que se ha presentado.

Agradezco a mis padres Ernesto Linarez y Carmen Arteaga de Linarez, por brindarme su apoyo incondicional, ofrecerme sus cuidados constantes y ser mi fuerza en los tiempos difíciles.

Agradezco a mi tía Violeta Linares por su apoyo ilimitado y la constante cooperación que siempre me brinda.

Agradezco a mi querido hermano Ernesto Linarez por su confianza depositada en mí para culminar este objetivo.

Agradezco al Dr. Wilmer Rafael Chávez Rea, quien me prestó en muchas ocasiones su ayuda para culminar este trabajo de grado.

Agradezco a mi tutor Prof. Ramón Valera, por orientarme y guiarme académicamente para culminar este objetivo propuesto.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

ÍNDICE DE ILUSTRACIONES

UNIVERSIDAD CENTROCCIDENTAL “LISANDRO ALVARADO”
DECANATO DE CIENCIAS Y TECNOLOGÌA
POSTGRADO EN CIENCIAS DE LA COMPUTACIÓN

PROPUESTA DE UNA ARQUITECTURA DE SOFTWARE PARA LA CONFORMACION DE UN REPOSITORIO DE DATOS DE REDES SOCIALES PARA ANÁLISIS Y MONITOREO

Autor(a): María Esperanza Linarez Arteaga.

Tutor(a): Ramón Valera.

RESUMEN

Las redes sociales a través de sus tecnologías han ayudado a la difusión y al intercambio de información de manera accesible y disponible al mayor número de personas. En los últimos años los usuarios han sentido la necesidad de saber con precisión que es lo que se dice en las redes sociales en relación a una temática, actor o ubicación geográfica, es por tal motivo que existen algunas propuestas a nivel de software que ya proporciona este tipo de seguimiento, sin embargo presentan algunas limitantes a la hora cubrir las necesidades demandadas por el usuario agregando a esto que los modelos existentes a nivel de arquitectura de software para cubrir estas soluciones son muy incompletos y a su vez la tecnología que usan es limitada. La siguiente investigación, tiene como propósito diseñar una arquitectura de software que garantice la conformación de un repositorio de datos provenientes de redes sociales para realizar procesos de análisis y monitoreo, dicho estudio se basa en una metodología de investigación documental bajo una modalidad de proyecto especial. El estudio quiere demostrar que logrando un repositorio consolidado de datos generados en redes sociales, para ejecutar técnicas de análisis y monitoreo, contribuya a mejorar estrategias de indexación que faciliten la ejecución de procesos de preparación y exploración de datos, cubierta por la minería de datos, con la finalidad de generar análisis exactos de la información almacenada, fomentado así la inteligencia de negocio en las organizaciones para mejorar el conocimiento de sus operaciones y las bases estratégicas para la toma de decisiones.

Palabras Claves: Redes sociales, Arquitectura de Software, Datos de redes sociales, repositorio de datos, análisis y monitoreo de datos, minería de datos

INTRODUCCION

Las redes sociales hoy en día son un medio importante para la comunicación, es por ende que empresas como facebook, Google, Linken In, entre otras, han innovado en tecnologías novedosas para que sus adeptos compartan información de manera rápida y confiable, generando así volúmenes de datos de manera constante y abundante, es por tal motivo que Villar (2011) asegura que las mismas se han convertido en un factor clave en la búsqueda de información.

La Comisión Europea (2010) asegura en su informe "Social Networks Overview: Current Trends and Research Challenges (Resumen de Redes Sociales: tendencias actuales y retos de investigación)", que el rápido crecimiento de popularidad de uso de redes sociales, ha originado nuevos desafíos a resolver, relacionados al manejo y gestión de la información generada en esos espacios .

Entonces es importante destacar que la información que se generan en redes sociales, es inmensa, diversa y dinámica, por lo que surge entonces, la necesidad por parte de los usuarios en la web de realizar medición de su entorno social, con la intención de conseguir información relevante como: ¿Qué se está diciendo de ellos mismos?,¿ Cuál es el tema de mayor interés del cual se está hablando? y ¿Como es su presencia o popularidad como usuario activo en la red social?.Es por esta razón que la Comisión Europea (2010) afirma que la información que se genera en redes sociales "es un campo emergente de investigación multidisciplinar" en donde se abordan los procesos que giran en torno al análisis de datos en redes sociales. Gundecha y Liu (2012) en su artículo " Mining Social Media: A Brief Introduction (Minería Social Media: Una breve introducción) " comentan que los datos que se generan en estos medios sociales no están fácilmente disponibles para la investigación científica, por lo tanto para facilitar su procesamiento los autores aseguran que se aplican técnicas y algoritmos de enfoque big data (datos grandes), atendiendo a que los contenidos generados en estas aplicaciones sociales tiene la propiedad de ser no estructurados.

Gundecha y Liu (2012) comentan que la búsqueda y recuperación de estos contenidos generados se realiza a través de métodos específicos, que en conjunto con técnicas big data logran gestionar la información heterogénea generada. Sin embargo el Instituto Nacional de Ciencia y Tecnología para la Web del Brasil (INWEB) y la Universidad Federal de Minas Gerais (UFMG) (2013) aseguran que cuando observan soluciones de software basadas en los procedimientos anteriores, se presentan dentro de la solución limitantes en cuanto al manejo de datos no estructurados para generar consultas más precisa, Chaudhuri (2012) asegura que esta problemática se debe a que no existe una infraestructura adecuada para la fase de preparación de los datos para análisis más profundos. Cabe destacar que la mayoría de estas soluciones no se apoyan en repositorios de datos no estructurados previamente almacenados, si no como lo resaltan Adedoyin (2013) en plataformas de agregación, que se enfocan en procesos de recopilación iterativa de información, que van hacia la fuente, localizan el dato y muestran el resultado al usuario, impidiendo así mantener un historial de datos generados por cada red social.

La conformación de un repositorio único de contenidos generados en redes sociales es una solución importante ya que sobre él se pueden aplicar procesos de exploración y monitoreo de datos, con el propósito de generar análisis posteriores más exactos en base a la información recolectada. Esto ayudaría a los investigadores a mejorar las técnicas ETL (Extraction, Transformation and Load) (Extracción, Transformación y Carga) y algoritmos de procesamiento utilizados en la minería de datos, que contribuiría a encontrar los patrones que se esconden en los enormes conjuntos de datos generados en estos espacios sociales y interpretarlos para el conocimiento.

Esta investigación propone una solución arquitectural de software con la finalidad de conformar un repositorio de datos provenientes de redes sociales para realizar procesos de análisis y monitoreo, con el objetivo que se pueda gestionar los distintos contenidos generados en redes sociales.

La investigación se circunscribe en una metodología de Proyecto Especial, apoyándose en una Investigación Documental, por lo tanto se revisan a fondo los conceptos y teorías necesarios para el desarrollo de la misma, el siguiente estudio se encuentra estructurado de la siguiente manera:

El Capítulo I, contempla el planteamiento del problema, objetivos general y específico, justificación, alcance.

Capítulo II, fundamentado el marco teórico, conformado por los antecedentes y bases teóricas en donde se establecen las teorías y experiencias anteriores inherentes al tema. En este punto se realizo un proceso de documentación que permitió conseguir avales teóricos y trabajos hechos con anterioridad para determinar el estado del arte.

Capítulo III, marco metodológico, en donde se explica la naturaleza de la investigación, diseño de investigación y métodos de recolección de información. Para dar repuesta a los objetivos planteados en la investigación, el diseño de la investigación estuvo enmarcado en un procedimiento conformado por tres (3) fases: la fase I, comprendió el diagnostico y análisis de la información recopilada, la fase II, oriento a determinar las características del repositorio, donde se almacenará los datos obtenidos de redes sociales y finalmente la fase III , corresponde al diseño de la arquitectura propuesta.

Capítulo IV, Análisis de los resultados, se detalla la propuesta del estudio, siguiendo las fases erigidas en la metodología, adicional se presentan los principales artefactos UML que describen las diversas vistas del sistema en estudio.

Capitulo V, Conclusiones y recomendaciones, es donde se plantean las conclusiones a las que se llegaron luego de la culminación del proyecto, así como también, las posibles recomendaciones para mejorar el mismo.

CAPITULO I

EL PROBLEMA

Planteamiento del Problema

En los últimos años las redes sociales han adquirido un poder de expansión como componente estratégico en el manejo de la información, esto se debe a la utilidad que han abarcado en las diferentes áreas productivas de la sociedad como fuente de información. Un ejemplo palpable de lo expuesto es que varias empresas del área de periodismo han tomado este tipo de herramienta como medio de comunicación pilar para obtener y transmitir información a sus clientes, tal como confirman Irragori y Cadíz (2011) en su programa de televisión Club de Prensa NT24 "es interesante ver el poder que están adquiriendo estos canales de comunicación como fuente rápida de información".

Por consiguiente el número de usuarios va en explosivo aumento, según el sitio web www.iredes.es en la presentación de su tercera versión del Mapa iRedes 2013 en el III Congreso Iberoamericano sobre Redes Sociales, Facebook registró más de 1.060 millones de usuarios siendo la más solicitada a nivel mundial seguida de Youtube con más de 800 millones de usuarios. En lo que respecta a Latinoamérica según la consultora comScore en su informe anual denominado "Futuro Digital Latinoamérica 2013", indica que en esta zona se invierten mensualmente 10 horas

visitando redes sociales, en donde se demostró que los usuarios latinoamericanos están altamente involucrados con estos medios sociales consumiendo casi 5 horas más al mes que el promedio global (véase la figura 1), siendo Facebook y Linkedln los sitios que obtienen más número de visitantes en la región. En el caso de Venezuela, con relación a redes sociales la empresa Socialbakers (citado por Amador, 2013) asegura que más de 9.579 200 millones de usuarios se encuentran registrados en Facebook , la cual es la red social con más usuarios en el país, la segunda como lo comenta Amador (2013) es twitter que cuenta con más de 6.400.000 millones de usuarios registrados.

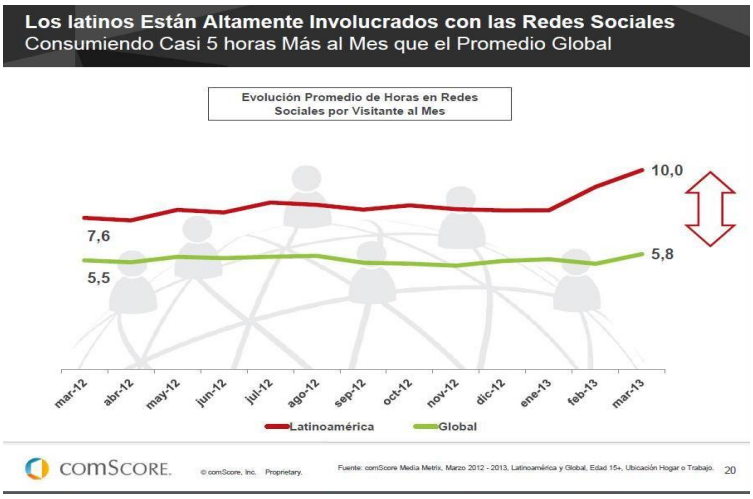


Figura 1. Evolución Promedio de Horas en Redes Sociales por Visitante al Mes.
Fuente: ComScore (2013)

Es evidente entonces que a través de las redes sociales millones de usuarios van generando volúmenes de datos de manera constante y abundante, como resultado de la información que comparten a través de los distintos mecanismos sociales, tales como blogs, foros, chats entre otros, originando así una enorme cantidad de información no estructurada. Estos datos generados en estos espacios sociales pueden representar a los usuarios un factor clave a la hora que ellos quisieran

interpretar, comprender y identificar lo que puede ser información y lo que simboliza tráfico, es por esta razón que hoy en día, tal como lo asegura Kaplan y otros (2010), varios usuarios sobre todo los corporativos, acuden a utilizar soluciones tecnológicas basadas en big data que estén orientadas a ejecutar procesos de análisis de datos de redes sociales en línea.

Gundecha y Liu (2012) mencionan que el análisis de datos de redes sociales en línea se apoyan en los principios de la minería de datos, la cual incluye procesos ETL como base para gestionar información de tipo big data, estos procesos como lo indica Martinez (2013) se caracterizan por ejecutar procedimientos de extracción de datos de diferentes aplicaciones, para su posterior transformación y siguiente carga, esta última fase se encarga de enviar los datos ya transformados en un formato deseado, a un sistema destino donde son procesados y analizados.

Adedoyin y otros (2013) asegura que cuando se trata de datos de redes sociales, la mayoría de las aplicaciones que trabajan con monitoreo y análisis de datos sociales se basan en plataformas de agregación, las cuales según el sitio de wikipedia.com, son sitios que recogen contenidos de múltiples servicios de redes sociales y los presentan de manera unificada, mostrando así la información recolectada en un solo lugar. Souravlias y otros (2012) comentan que este tipo de herramienta genera resultados en tiempo real para los usuarios, no obstante en esta aplicaciones se observan limitantes como la ausencia de búsquedas parametrizadas y la visualización de un histórico de datos. Chulis (2013) explica que lo anterior ocurre debido a que en este tipo de soluciones dejan un poco de lado la infraestructura de almacenamiento como base para ejecutar análisis a posterior, por lo que la autora afirma que "si la información no se recolecta o se almacena adecuadamente y periódicamente, los datos de medios sociales son básicamente efímeros" Chulis (2013).

Souravlias y otros (2012) comentan que existen algunas soluciones que se apoyan en repositorios de datos de redes sociales para obtener un análisis más completo. Gundecha y Liu (2012) describen que los datos que se encuentran

almacenados en estos repositorios presentan la característica de ser no estructurados, sin embargo Chaudhuri (2012), asegura que cuando se trata de recuperar nuevamente la información almacenada para aplicar procesos de análisis y monitoreo, se presentan nuevos desafíos que todavía existen por resolver, entre los retos que formula Chaudhuri (2012) están: ¿Cómo identificar fácilmente los fragmentos relevantes de los datos que provenga de una fuente específica?, ¿Cómo utilizar técnicas de limpieza de datos para fuentes de datos parecidas? y ¿Cómo ver los resultados progresivamente a través de una consulta?

Gundecha y Liu (2012) indican que construir una solución que pueda dar repuestas de manera simultánea a las interrogantes anteriores, es complicado y es debido a que los datos que se generan en redes sociales son enormes, ruidosos, distribuidos, no estructurados y dinámicos. Por otra parte existe el dilema en cuanto a un modelo conceptual inicial a seguir, debido a que cuando se trabaja con datos grandes es muy inexacto seguir un patrón que pueda ayudar a modelar una solución que nos indique con detalle aspectos como: identificar los elementos necesarios que intervienen en la construcción de un repositorio de datos no estructurados, mecanismos apropiados de búsqueda que se pueda aplicar en estos datos ya almacenados y las características especificas que presenta estos depósitos de información no estructurada. Chisholm (2012) menciona que las restricciones anteriores se debe a que en big data no hay una aproximación del modelo conceptual y el modelo de datos lógico como se observa en el paradigma relacional.

Los mecanismos de recuperación de estos datos no estructurados, se basan en métodos que han sido implementados en algunos estudios de análisis de redes sociales, uno de ellos es a través de las aplicaciones ad-hoc, Canali, Colajanni y Lancellotti (2011) indica que los ad-hoc son aplicaciones de terceros que explota un conjunto de APIs, sin embargo Nazir (2008) (citado por Canali, Colajanni y Lancellotti , 2011) comenta que el uso de aplicaciones ad-hoc para adquirir datos de redes sociales no es optimo, debido a que siempre hay que esperar la autorización por parte del usuario, lo que origina que el conjunto de datos disponible para el

análisis resulta limitada e inútil. La otra técnica mencionada por la literatura es el crawler (rastreo), tal como lo define Canali, Colajanni y Lancellotti (2011) es una búsqueda que se ejecuta en forma iterativa hacia las diferentes fuentes sociales, apoyándose de APIs públicas que ofrecen algunos operadores de red social.

Actualmente existen soluciones a nivel arquitectural en donde se enfocan en resolver la obtención y manipulación de los datos generados en redes sociales, algunas de estas arquitecturas están basadas en componentes mientras que otras están basadas en agentes, en el caso de esta última se han definido agentes abstractos para implementar el proceso de rastreo por cada tipo de dato a procesar (documentos, imágenes, videos, audios y otros), como lo detalla Ghaderi y otros (2010) en su trabajo denominado "A Social Network-based Meta Search Engine (Metabuscador basado en redes sociales)", sin embargo en este estudio se demostró que al ejecutar las primeras iteraciones del modelo en tiempo real, notaron que la instanciación del agente rastreador es muy compleja, debido a los diferentes intervalos de tiempo de rastreo utilizados por cada tipo de dato. Con respecto a las soluciones basadas en arquitectura de componentes, el proceso de rastreo de datos es parecido al de arquitectura basada en agentes, con la diferencia de que el componente rastreador al obtener el dato lo envían a un segundo componente que ejecuta el proceso de transformación y finalmente lo envía a un repositorio de datos no estructurados.

Canali, Colajanni y Lancellotti (2011) proponen que para que el proceso de extracción de datos se ejecute de forma distribuida y así cubrir todas las fuentes sociales existente en la web, el componente rastreador se debe ejecutar en un servicio en la nube, sin embargo Canali, Colajanni y Lancellotti (2011) resaltan que esta tecnología restringe en un periodo de tiempo específico el número de peticiones permitidas, impuesto por el operador de la red social, lo que trae como consecuencia que el conjunto de datos devuelto por la API es incompleto, originando así que la información obtenida no es lo suficiente para explorar la estructura de la red social.

Por otra parte, las soluciones que se basan en plataforma de agregación para ejecutar análisis de datos de redes sociales, utilizan el método de agregación de contenidos, que es la recopilación de contenidos que se generan en múltiples fuentes en línea, enfocados en un tema en especifico, generando así las salidas de información solicitada por el usuario, pero sin manejar ningún tipo de almacenamiento de datos, es por tal motivo que en la web se observa aplicaciones de este tipo como: Socialmention, ReputaciónXL, Buzzmonitor, Kurrently, entre otras, que no ofrecen consultas con parámetros específicos, ni visualización de un histórico de datos previamente consultados, lo que impide al usuario distinguir el comportamiento que ha tenido cierto ítem información en el tiempo. Otra limitante que se presenta en este tipo de soluciones, es en el manejo de un solo tipo de dato, como es el caso de FacebookLexicon que simplemente se enfoca en buscar las conversaciones emitidas y no en los otros contenidos que están disponibles en la aplicación Facebook.

Es claro entonces que las limitantes mencionadas se deben a que estas aplicaciones se basan en soluciones arquitecturales en donde no contemplan el manejo de un repositorio de datos no estructurados previamente almacenados, por lo tanto es complicado bajo la circunstancia mencionada, aplicar de manera uniforme métodos de recuperación y gestión para los distintos tipos de datos generados en redes sociales, por consiguiente esto limita la ejecución de procesos de análisis y monitoreo de información que se genera en las redes sociales. Por lo anterior expuesto Chaudhuri (2012) menciona que es necesaria la construcción de un entorno que permita la exploración datos no estructurados para análisis más profundos.

Por lo tanto, se hace necesario el estudio de una solución a nivel de arquitectura de software que esté basada en la obtención de datos provenientes de las diferentes redes sociales y proceder a almacenarlos con el fin de generar un repositorio de datos no estructurados que sirva para aplicar procesos de análisis y monitoreo, esto permitirá facilitar el seguimiento a largo plazo de los datos o contenidos generados en una red social particular.

Ante este planteamiento, surge la necesidad de formular las siguientes interrogantes:

¿Cuáles son los elementos necesarios para un proceso de construcción, alimentación y extracción de un repositorio de datos en el marco de las redes sociales?

¿Cuáles son los posibles mecanismos adecuados de búsqueda y recuperación de contenidos de diferentes redes sociales?

¿Cuáles serían las características de un repositorio de datos no estructurados que permitan almacenar datos de redes sociales para realizar procesos posteriores de análisis y monitoreo?

¿Cuál es la arquitectura de software adecuada que permita la obtención y almacenamiento de información de tipo no estructurada generada en las distintas redes sociales?

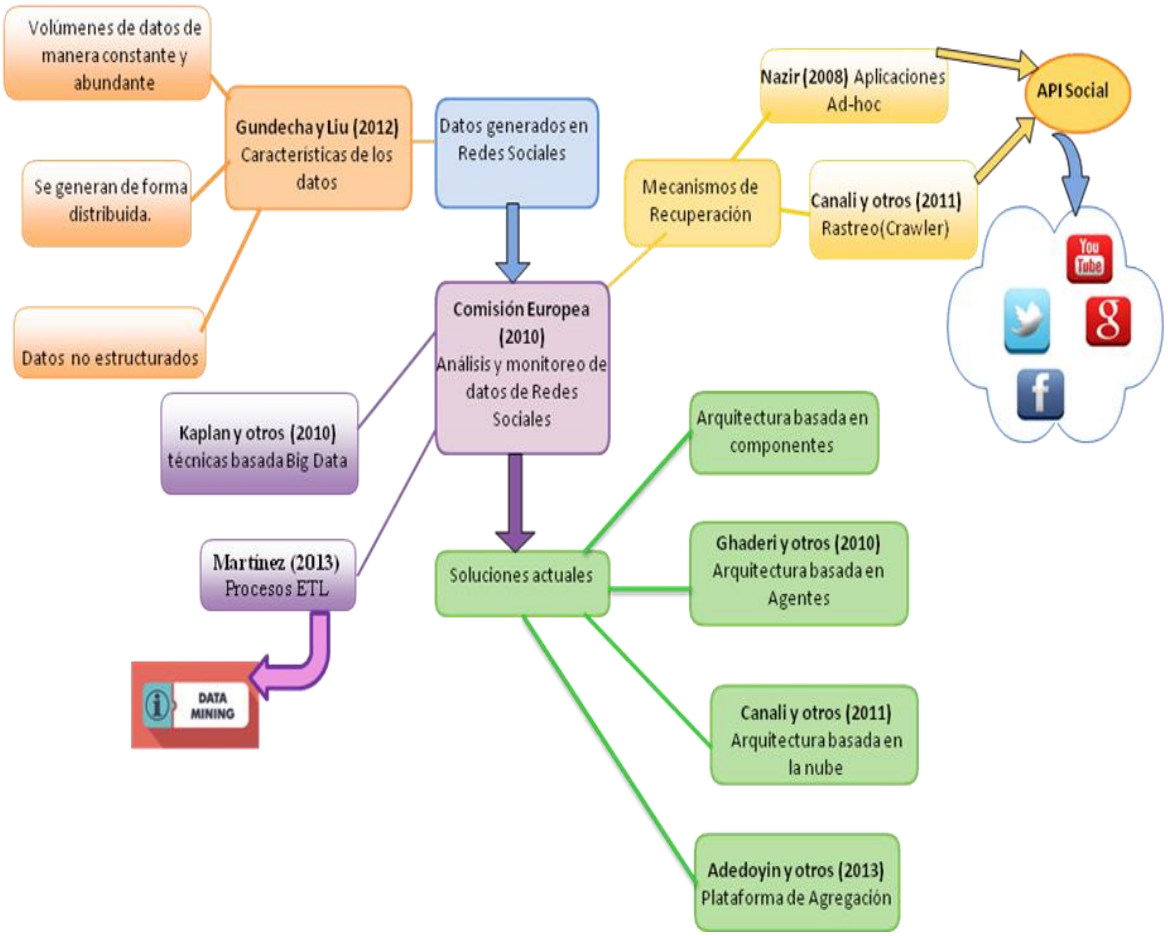


Figura 2.Mapa conceptual del Contexto del Problema.
Fuente: Autor de la Investigación

Objetivos de la Investigación

Objetivo General

Proponer una arquitectura de software para conformar un repositorio de datos provenientes de redes sociales para realizar procesos de análisis y monitoreo.

Objetivos Específicos

1. Identificar los elementos necesarios al proceso de construcción, alimentación y extracción de un repositorio de datos en el marco de redes sociales, así como también los mecanismos de búsqueda y recuperación de los diferentes contenidos generado en las redes sociales.
2. Determinar las características de un repositorio de datos no estructurados que permita almacenar datos de redes sociales para realizar procesos posteriores de análisis y monitoreo.
3. Diseñar la arquitectura de software adecuada que permita la obtención y almacenamiento de información de tipo no estructurada generada en las distintas redes sociales.

Justificación e Importancia

Las redes sociales tienen una importancia significativa en la creación de información, Kaplan y Haenlein (2010) aseguran que las mismas han permitido que la difusión y el intercambio de datos estén disponibles de manera accesible y rápida para todos los usuarios en la web, pero como lo indica Kaushik (2011), el aumento explosivo de la información generada allí, genera la inquietud por parte de algunos usuarios e investigadores, la necesidad de identificar los datos que realmente representen valor para traducirlos en conocimientos.

El auge de las redes sociales ha hecho imprescindible la creación de soluciones tecnológicas que midan los resultados de las acciones, que los usuarios realizan en sus perfiles sociales y determinen si el impacto y los efectos conseguidos son los deseados. Esto ha permitido entonces identificar y estudiar los procesos que se encuentren involucrados en la obtención y análisis de los contenidos expuestos en las redes sociales.

La propuesta de una arquitectura de software para conformar un repositorio de datos provenientes de redes sociales que sirva para realizar procesos de análisis y monitoreo, es el punto de partida de tener cierta independencia de ejecutar análisis a posterior más exactos a partir de la información recolectada. Esto ayudaría a los investigadores en mejorar y diseñar nuevos algoritmos de procesamientos que apoyen a optimizar los procesos ETL utilizados en la minería de datos, los cuales aportarían a la construcción de mecanismos de medición más precisos que ayuden a encontrar los patrones que se escondan en los enormes conjuntos de datos heterogéneos, con el fin de interpretarlos para el conocimiento y convertirlos en información útil.

Por lo tanto, se espera que el presente estudio sea de importancia a nivel académico, ya que presenta una referencia inicial del modelaje bajo enfoque big data, del cual Michael y Miller (2013) confirman que aunque es un tema un poco abstracto de abordar pero que en los últimos años está tomando más forma y consistencia, debido a la alta demanda de analizar y gestionar los distintos tipos de datos que se generan en las distintas fuentes de información, por tal motivo se espera que el actual trabajo pueda ser utilizado como investigación preliminar para realizar otros estudios similares que ayuden a organismos, universidades e instituciones educativas o públicas a obtener una mejor precisión de la información generada en sus entornos sociales con la finalidad de mejorar el conocimiento de sus operaciones que ayudarían a los procesos de toma de decisiones a cumplir los objetivos planteados por la organización.

Por otra parte, la investigación puede aportar información relevante sobre la temática planteada, convirtiéndose en una ayuda para otros investigadores que

trabajen con línea investigativa de minería de datos y operaciones con datos grandes, fortaleciendo así la literatura existente sobre el análisis de los datos de las redes sociales en línea.

Alcances y Limitaciones

La presente investigación, está enmarcada en la propuesta de una arquitectura de software con sus respectivos componentes que garantice la obtención y el almacenamiento de los contenidos generados en redes sociales, dando escalabilidad y reusabilidad. El estudio no contempla la puesta en marcha en un entorno de producción de lo anterior descrito.

Esta investigación se centra en la obtención de información generada por los distintos medios sociales, basándose en procesos que permitan el almacenamiento de los datos, con el fin de conformar un repositorio que pueda dar soporte a procesos de análisis y monitoreo. El enfoque propuesto se basará en definir procesos que ayuden a la obtención y transformación de los distintos tipos de datos provenientes de las diferentes aplicaciones de red social, para luego su posterior almacenamiento que ayudará a garantizar su indexación para posteriores consultas.

La investigación no tratará el área de minería de datos que aborda los procesos orientados a identificar el grado de sentimiento (negativo, positivo, neutral) de la información almacenada, está problemática puede ser tratada como tema aparte o complementario en una futura investigación.

En lo que respecta a las limitaciones, se observa el nivel de privacidad que presenta los perfiles sociales, lo cual esto limita la zona de exploración que tenga el proceso de rastreo para efectuar la extracción de la información.

La solución va estar sujeta al nivel de seguridad que autorice el usuario para mostrar información, es decir los componentes encargados de ejecutar el proceso de

rastreo y extracción, van a llegar hasta el nivel de acceso que autorice la API social correspondiente.

CAPITULO II

MARCO TEORICO

Antecedentes de la Investigación

Son muy variadas las investigaciones realizadas donde se observan propuestas basadas en arquitectura para resolver el análisis de información generada en las redes sociales, así como también técnicas que se emplean para tratar estos contenidos para luego ser almacenados. En esta sección se hará mención de aquellos trabajos que han antecedido a la presente investigación, que de alguna manera u otra ayudan a la misma y sirven de base para la solución del problema planteado.

Como primer antecedente tenemos el de Canali y otros (2011), en su artículo denominado "Data Acquisition in Social Networks: Issues and Proposals" (Adquisición de datos en redes sociales: Problemas y Propuestas), describen de manera breve las tres principales técnicas propuestas por la literatura para adquirir datos generados en redes sociales: network traffic analysis (análisis del tráfico de red), ad-hoc applications(aplicaciones ad-hoc) y crawling the user graph (rastreo del grafo de usuario), enfatizando que los dos últimos métodos son lo más usados.

Posteriormente proponen una solución para mejorar la adquisición de datos de redes sociales utilizando el enfoque cloud computing (computación en la nube), en donde destacan el funcionamiento de técnicas de crawlers que trabajan de manera paralela a través de maquinas virtuales. El Crawler implementado en esta solución está basado en una arquitectura modular conformada por dos partes principales: La

primera parte contiene el motor responsable de la coordinación de la ejecución paralela de los Crawlers. La segunda parte es el módulo de recolección de datos, (ver figura 3) el cual interactúan con el motor de coordinación para la adquisición de datos relacionado con cada red social utilizando sus APIs correspondientes. Este trabajo sirvió de insumo, ya que permitió conocer el estado del arte, acerca de los métodos de recuperación existentes que se utilizan para acceder a estos datos, comprobando de manera detallada que se puede utilizar algunas de estas técnicas en forma combinada a través de una solución arquitectural.

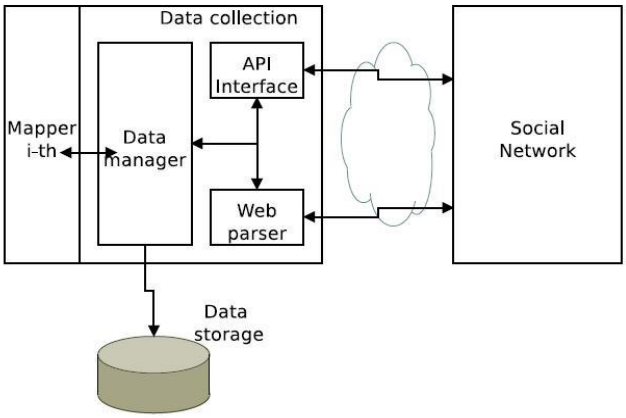


Figura 3 Módulo de recolección de datos.
Fuente: Canali y otros (2011)

Continuando con los estudios que detallan técnicas de recolección y recuperación de datos en redes sociales tenemos el trabajo realizado por Catanese y otros (2011) en su artículo denominado “Crawling Facebook for Social Network Analysis Purposes" (Rastreo de Facebook para fines de análisis de redes sociales), presentan una solución referente a la recolección masiva y análisis de datos que generan las conexiones entre los participantes pertenecientes a una red social, allí describen enfoques alternativos para la recopilación y manipulación de datos de red social, los cuales fueron implementados en una red social especifica, como fue en este caso el sitio Facebook. Este artículo se toma como marco de referencia para

entender las tareas que se encuentran involucradas, en un proceso de recolección de datos en redes sociales (ver figura 4), el cual sirvió como base para proponer las acciones básicas que ejecutaran los componentes de extracción y gestión de datos de redes sociales que estarán habilitados en la arquitectura.

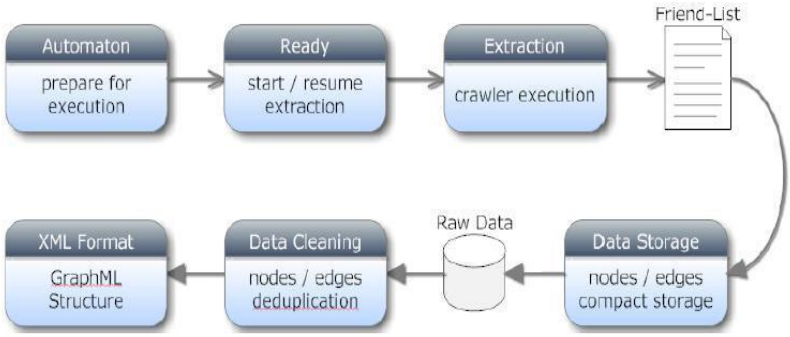


Figura 4 Diagrama de estado del proceso de minería de datos.
Fuente: Catanese y otros (2011)

Por otra parte, Boukhanovsky y otros (2011) en su artículo denominado “A generic Architecture for a Social Network Monitoring and Analysis System" (Una arquitectura genérica para monitorear y analizar una red social), define una arquitectura genérica de software que facilita el seguimiento a largo plazo de diversas redes sociales existentes y emergentes. La arquitectura planteada consta de tres módulos: el crawler, el repositorio y el analizador (Ver figura 5). Cada módulo presenta un conjunto de componentes que realiza tareas específicas y la comunicación entre ellos es a través de interfaces propias que cada modulo posee. El trabajo presentado por estos autores, sirve como aporte a la presente investigación, debido a que se tomó como base la guía arquitectural y el modelo de requisitos, que permiten observar las características mínimas que deben estar presentar en una arquitectura que maneje datos provenientes de las distintas redes sociales y luego gestionar su almacenamiento en un repositorio único.

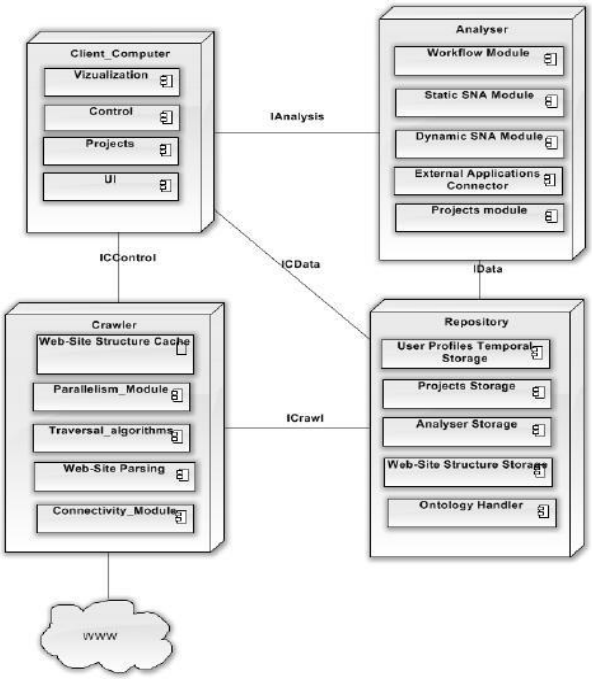


Figura 5. Arquitectura genérica para el Monitoreo y Análisis de una Red Social.
Fuente: Boukhanovsky y otros (2011)

Continuando con los estudios que describen modelos arquitecturales para manejar datos de distintas redes sociales, se observa los trabajos presentados por Mysore y otros (2013), el primero denominado "Big data architecture and patterns, Part 4: Understanding atomic and composite patterns for bigdata solutions" (Arquitectura Big Data y patrones, parte 4: Entendiendo patrones atómicos y compuestos para soluciones Big Data) y el segundo denominado "Big data architecture and patterns, Part 5: Apply asolution pattern to your big data problem and choosethe products to implement it" (Arquitectura Big Data y patrones, parte 5: aplicar un patrón de solución a su problema Big Data y elegir productos para implementarlo). Ambos artículos explica una serie de pasos y patrones que representan una orientación a los usuarios expertos para diseñar y aplicar una solucion big data dependiendo de los requerimientos identificados en su organizacion,

las tecnicas detalladas en la investigacion reflejan a través de pasos y procesos que han sido observados en la mayorias de la soluciones de software que gestionan con datos de tipo no estructurados, estructurados y semi-estructurado que son generados en diversas fuentes de informacion como social media, dispositivos digitales, lenguaje de humano y otros. El estudio aporta a la investigación un marco de referencia para entender un modelo arquitectural (ver figura 6), basado en el enfoque big data, en donde plantea un modelo que describe los elementos importantes que intervienen en el manejo de los datos no estructurados que se generan en redes sociales. Cabe destacar que el estudio presenta una documentación detalla acerca de los procesos que intervienen en la gestión de datos no estructurados.

Siguiendo con los estudios tenemos el trabajo realizado por Souravlias y otros (2012), llamado “InterSocialDB: An Infrastructure for Managing Social Data" (InterSocialDB: una infraestructura de gestión de datos de contenido social), en donde proponen una arquitectura para almacenar y analizar datos provenientes de aplicaciones de redes sociales. Dicha arquitectura está conformada por dos componentes: el componente de adquisición de Datos y el componente de procesamiento de datos (ver figura 7). Este trabajo sirvió de insumo para verificar otros fundamentos que se deben tomar en cuenta al plantear los procesos de recolección de datos de redes sociales, adicional se tomó como base las alternativas descritas allí que se usan para definir un almacenamiento para datos de tipo no estructurados, adicional se evaluó el modelo arquitectural propuesto, para reforzar el funcionamiento de los componentes que interactúan en la arquitectura propuesta.

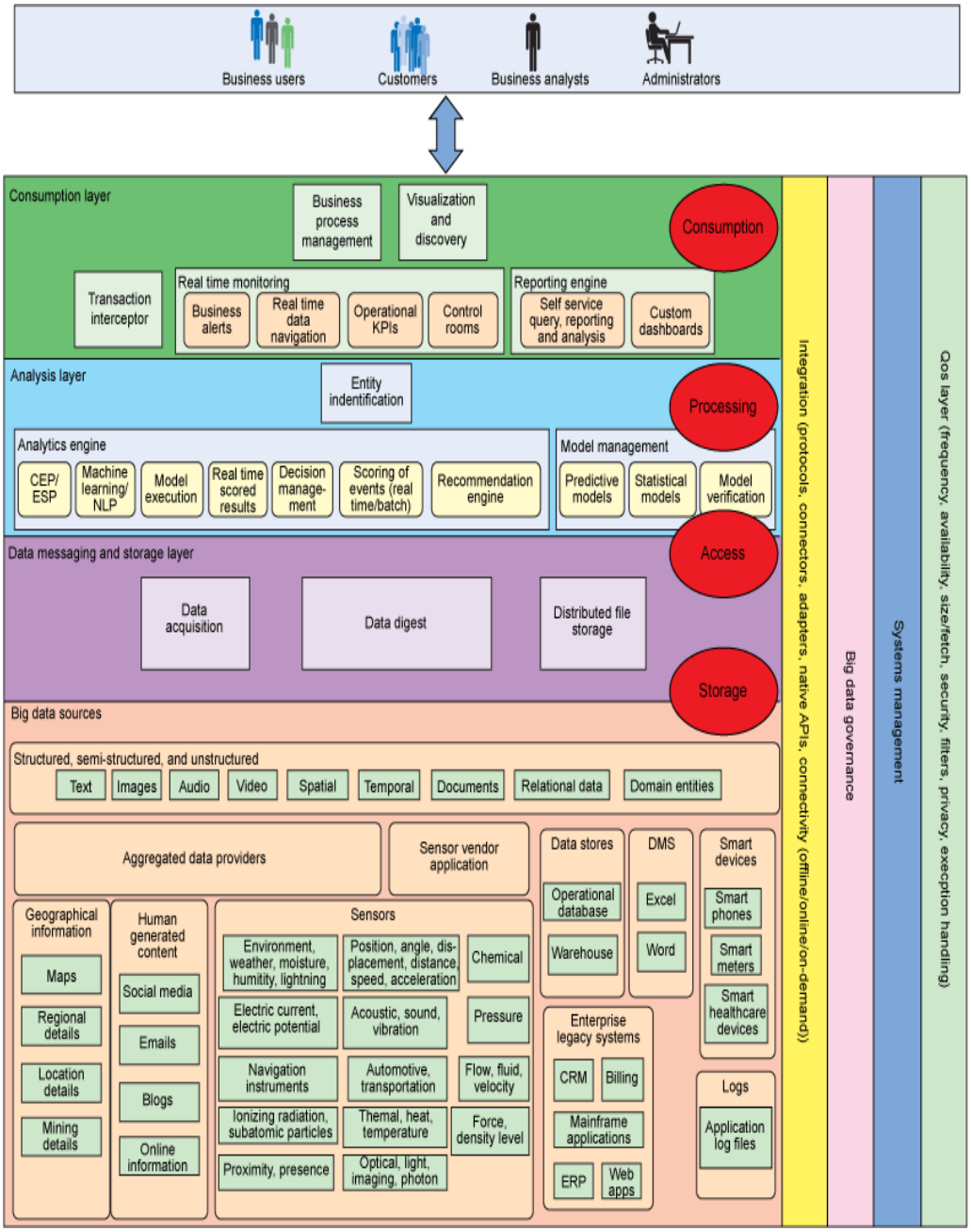


Figura 6. Arquitectura genérica para soluciones Big Data.
Fuente: Mysore y otros (2013)

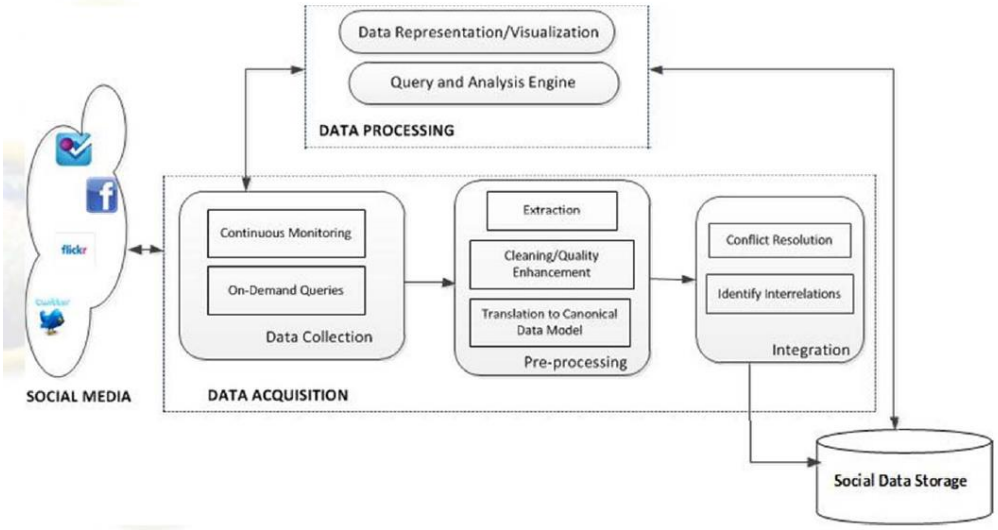


Figura 7. InterSocialDB: Una infraestructura de gestión de datos sociales.

Fuente: Souravlias y otros (2012)

Con respecto a estudios relacionados al almacenamiento de contenido generados en redes sociales podemos observar el Proyecto de Final de carrera realizado por Morros (2013) denominado "Big Data- Análisis de Herramientas y Soluciones", en donde hace un estudio dentro de un marco teórico de las distintas herramientas utilizadas en el área de big data, resaltando el uso de bases de datos NoSQL como plataforma adecuada para soportar grandes volúmenes de datos. El aporte a esta investigación radica en la definición y clasificación acercas de las diferentes técnicas existentes para tratar el almacenamiento de data heterogénea proveniente de redes sociales.

Por último tenemos el trabajo de Martínez (2013), denominado "Desarrollo de una Herramienta de Inteligencia de Negocio para el análisis de redes sociales almacenada en grafos", el cual consistió en el desarrollo de una herramienta Bussines Intelligence (Inteligencia de Negocios) que permitiera la definición y el cálculo de indicadores de interés para analizar redes sociales. Este artículo se toma como marco

de referencia para entender el conjunto de datos de redes sociales y el almacenamiento en un gestor de base de datos orientado a grafo, en donde a través de técnicas NoSql, describen las opciones que se implementan para poder realizar ciertas consultas en dicha base de datos.

Con lo anteriormente expuesto se puede concluir que existen trabajos donde se han empezado a profundizar en el conocimiento referente a la obtención, tratamiento y almacenamiento de información que se generan en las diferentes aplicaciones de red social, esto ha originado que los investigadores realicen estudios orientados a propuestas arquitecturales existentes que podrían ayudar a generar nuevos procesos de valor agregado a técnicas o herramientas de medición, las cuales estén orientadas a gestionar los diferentes contenidos generados en redes sociales. Basada en estas experiencias y en vista al éxito logrado en la definición de metodologías de extracción de datos generados en redes sociales, se plantea una propuesta de arquitectura de software para la conformación de un repositorio de datos de redes sociales para análisis y monitoreo.

En el cuadro 1, se observa un resumen de los antecedentes planteados en esta sección, se describen los objetivos y aportes de cada antecedente:

Cuadro 1
Resumen de Antecedentes

Antecedente	Objetivo	Aporte
Canali y otros(2011) *"Data Acquisition in Social Networks: Issues and Proposals"*	Describen de manera breve las tres principales técnicas propuestas en la literatura para adquirir datos generados en redes sociales: Análisis de trafico de red, Aplicaciones Ad-hoc, Rastreo del grafo del usuario	Contribuye a la documentación acerca de métodos de recuperación existente, que se utiliza para obtener los datos generados en las diferentes redes sociales.

Catanese y otros(2011) *Rastreo de Facebook para fines de análisis de redes sociales*	Definir una solución para efectuar recolección masiva y análisis de datos de las conexiones entre los participantes de redes sociales.	Presenta de manera esquematizada el proceso de tareas que implementaron para la recolección de datos en redes sociales, el cual servirá como base para proponer las acciones básicas que ejecutaran los componentes de rastreo y extracción de datos, que estarán habilitados en la arquitectura.
Boukhanovsky A , Semenov A, Veijalainen J. (2011) *Una arquitectura genérica para Monitorear y Analizar una Red Social*	Define una arquitectura genérica de software que facilita el seguimiento a largo plazo de diversas redes sociales existentes y emergentes	Modelo Arquitectural y Modelo de requisito
Mysore, Khupat y Jain (2013) *Arquitectura Big Data y patrones, parte 4: Entendiendo patrones atómicos y compuestos para soluciones Big Data*	Define una serie de pasos y patrones utilizados que representan una guía a los usuarios expertos para implementar una solución big data en el ámbito organizacional, el estudio aborda los problemas más comunes y recurrentes que se presenta en las soluciones de datos grandes. Este tutorial ayuda a identificar paso a paso los componentes necesarios que puedan conformar una arquitectura big data.	Plantilla arquitectural para modelar una solución big data. Componentes básicos que deben existir para conformar una arquitectura big data.
Mysore, Khupat y Jain (2013) *Arquitectura Big Data y patrones, parte 5: aplicar un patrón de solución a su problema Big Data y elegir productos para implementarlo*	El artículo describe los patrones de solución que pueden ayudar a definir una solución datos grandes(big data), usando un enfoque basado en escenarios. Cada escenario comienza con un problema de negocio y la descripción del por qué es necesaria la solución big data. Al	Modelo Arquitectural Documentación detallada acerca, de los procesos que intervienen en la gestion de datos no estructurados que se generan en redes

	final del artículo, se sugieren algunas herramientas y productos típicos orientados a IBM.	sociales, estos son: captura, transformación y almacenamiento.
Morros (2013) *Big Data- Análisis de Herramientas y Soluciones*	estudio que demuestra dentro de un marco teórico las distintas herramientas utilizadas en el área de Big Data	Demuestra la definición y clasificación de las diferentes técnicas existentes, para tratar el almacenamiento de dato no estructurados proveniente de redes sociales
Souravlias et (2012) *InterSocialDB:Una infraestructura para gestionar Datos sociales*	Propone una arquitectura para almacenar y analizar datos provenientes de aplicaciones de redes sociales.	Demuestra a través de su modelo arquitectural los procesos mínimos que deben existir para la recolección y almacenamiento de datos no estructurado generados en redes sociales, adicional se toma como base las alternativas descritas allí que se puede emplear para definir un repositorio de datos de tipo no estructurado.
Martínez (2013) *Desarrollo de una Herramienta de Inteligencia de Negocio para el análisis de redes sociales almacenada en grafos*	Desarrollo de una herramienta de inteligencia de negocio que permite la definición y el cálculo de indicadores de interés para analizar redes sociales.	Se basa en la utilización de gestores de base de datos orientados a grafos, para el almacenamiento del grafo de la red social con sus respectivos nodos, con la finalidad de mantener el historial de los perfiles de usuarios ya visitados

Fuente: Autor de la Investigación

Bases Teóricas

En esta sección se exponen las bases teóricas relacionadas con la investigación que se está desarrollando, donde se abarcan los diferentes conceptos claves para que el lector pueda captar de forma rápida la idea principal de la misma.

La investigación contempla los siguientes conceptos principales que se mencionarán a lo largo de la narrativa y corresponden a la Social Media, Redes Sociales, Herramientas de Monitoreo social, APIs, APIs Social, Arquitectura de Software, Componente de Software, Crawler, Contenido Generado por Usuario (UGC), Repositorio de Datos, Procesos ETL, Minería de Datos, Big Data, Arquitectura Big Data, NoSql, Técnicas de Modelado en NoSQL, Plataforma de Agregación.

Social Media

El social media, es una definición que ha venido evolucionado, el cual agrupa conceptos y herramientas que fusionan comunicación, publicación, y divulgación de contenidos de multimedia, el cual muchas veces es confundida por las herramientas y aplicaciones que lo llevan a cabo.

Kaplan y otros (2010), plantea que el social media es un grupo de aplicaciones basadas en internet que se basan en los fundamentos ideológicos y tecnológicos de Web 2.0 y que permiten la creación y el intercambio contenido generado por usuario.

La dimensión social de estos medios según Gilly (2003) (citado por Kaplan y otros 2010) está en el hecho de que el objetivo es de influir en los demás para ganar recompensas, por otro lado el individuo es conducido a crear una imagen que es consistente con la identidad personal, Gilly (2003) continua comentando que la

creciente plataforma tecnológica hace que el individuo desea expandirse por el ciberespacio.

Redes Sociales

De acuerdo con Boyd y Ellison (2007) (citado por Flores 2009), una red social se define como un servicio que permite a los individuos construir un perfil público o semi-público dentro de un sistema delimitado.

Flores (2009) resalta que un punto importante a tener en cuenta en las redes sociales es el término "efectos de red" que hace referencia al valor de una red con respecto al crecimiento de sus usuarios. Flores (2009) denomina "efecto de red" al tipo particular de externalidad que se produce cuando cada nuevo usuario añade valor a un producto por el hecho de unirse a la comunidad de usuarios, es decir cuantos más miembros tiene la red de usuarios más valor tiene para un miembro pertenecer a ella.

O'Reilly (2005) (citado por Flores 2009) asegura que un elemento esencial para crear una red exitosa y diseñar una arquitectura de participación, es en establecer las preferencias de los usuarios para compartir contenidos en forma automática, de modo que los usuarios contribuyan al valor de la red.

Debido a que las redes sociales son sitios que propician la interacción de miles de personas en tiempo real, el cuadro 2 presenta la clasificación realizada por Burgueño (2009) basado en la utilidad que se le da a las mismas:

Cuadro 2
Clasificación de la Redes Sociales

Clasificación	Tipo de Redes Sociales	Características	Ejemplo
Por Público Objetivo y Temática	Redes sociales horizontales	Son aquellas dirigidas a todo tipo de usuario y sin una temática definida.	Facebook, Twitter, Orkut
	Redes sociales verticales	Están concebidas sobre la base de un eje temático agregador. Su objetivo es el de congregar en torno a una temática definida a un colectivo concreto. En función de su especialización, pueden clasificarse a su vez en: ✓ Redes sociales verticales profesionales. ✓ Redes sociales verticales de ocio. ✓ Redes sociales verticales mixtas.	Xing , Linked In, Wipley, Minube Dogster
Por el Sujeto Principal de la Relación	Redes sociales humanas	Son aquellas que centran su atención en fomentar las relaciones entre personas uniendo individuos según su perfil social y en función de sus gustos, aficiones, lugares de trabajo, viajes y actividades	Koornk, Dopplr, Youare y Tuenti.
	Redes sociales de contenidos	Las relaciones se desarrolla uniendo perfiles a través de contenido publicado, los objetos que posee el usuario o los archivos que se encuentran en su ordenador	Scribd, Flickr, Bebo, Friendster
	Redes sociales	Conforman un sector novedoso entre las redes sociales. Su objeto es unir marcas,	Respectance.

	de inertes	automóviles y lugares. Entre estas redes sociales destacan las de difuntos, siendo éstos los sujetos principales de la red.	
Por Localización Geográfica	Redes sociales sedentarias	Este tipo de red social muta en función de las relaciones entre personas, los contenidos compartidos o los eventos creados.	Rejaw, Blogger, Kwippy, Plaxo, Bitacoras.com, Plurk
	Redes sociales nómadas	A las características propias de las redes sociales sedentarias se le suma un nuevo factor de mutación o desarrollo basado en la localización geográfica del sujeto. Este tipo de redes se componen y recomponen a tenor de los sujetos que se hallen geográficamente cerca del lugar en el que se encuentra el usuario, los lugares que ha visitado o aquellos a los que tenga previsto acudir.	Latitud, Brigthkite, Fire Eagle y Scout.
Por Plataforma	Red social MMORPG y metaversos	Normalmente construidos sobre una base técnica Cliente-Servidor	WOW, SecondLife, Lineage
	Red social web	Su plataforma de desarrollo está basada en una estructura típica de web.	MySpace, Friendfeed y Hi5

Fuente: Tomado de Burgueño (2009)

Herramientas de Monitoreo Social

Según la pagina web el sitio NeoHumano (2011), son herramientas que consiste en rastrear y procesar información sobre un tema o marca en redes sociales; en base a objetivos claros y métricas que permitan evaluar resultados.

En el enfoque de monitoreo de social media, se contempla tres fases principales que contempla la buena utilización de estas herramientas: identificar, rastrear y generar.

Kaplan y otros (2010) detallan que en la fase de identificación el individuo encuentra los temas, los usuarios y las plataformas relevantes, en la fase de rastreo el usuario localiza contenido propio y conversaciones sobre otros contenidos y la última fase denominada generar es la capacidad de respuesta que presenta el usuario para interactuar y elaborar estrategias para la toma de decisiones acertadas.

APIs

API (Application Program Interface). Conjunto de funciones y procedimientos que provee un sistema operativo, una aplicación o una biblioteca que definen de cómo pueden invocarlo una determinada función de un programa desde una aplicación. Cuando se intenta estandarizar una plataforma, se estipulan unos APIs comunes a los que deben ajustarse todos los desarrolladores de aplicaciones.

Arquitectura de Software

Clements (1998) (citado por Valera 2010), donde describe que la Arquitectura de Software es a grandes rasgos, una vista del sistema que incluye los componentes principales del mismo, la conducta de esos componentes según se la percibe desde el resto del sistema y las formas en que los componentes interactúan y se coordinan para alcanzar la misión del sistema.

Valera (2010) resalta que todas las definiciones de Arquitectura de Software radican en que es una estructura del sistema, representada o descrita a través de sus componentes y como estos se relacionan entre sí. Para representar esta estructura se cuenta con diagramas y vistas para representar las distintas perspectivas del sistema a desarrollar, al análisis de todo en conjunto permite al equipo de desarrollo una visión holística del sistema que representa.

La importancia de una buena arquitectura es que nos ayuda a obtener una mejor abstracción del sistema, organizar su desarrollo y presentar un lenguaje común para el equipo de trabajo involucrado (diseñadores, desarrolladores y usuarios). Adicional aporta una productividad en el desarrollo a través reutilización de componentes y monitoreo de otras áreas del ciclo de vida del software que permite resolver problemas como de portabilidad, escalabilidad, seguridad, entre otros

Modelos Arquitecturales

Pressman y otros (2006) (citado por Gómez 2011) El diseño arquitectónico se puede representar mediante uno o más modelos diferentes, en el Cuadro 3 se puede detallar la siguiente clasificación:

Cuadro 3
Modelos Arquitectónicos de Software

Modelo Arquitectónico	Descripción
Estructurales	Representan la arquitectura como una colección organizada de componentes de programa.
Marco de trabajo	Aumentan el nivel de abstracción del diseño en un intento de identificar los marcos de trabajo (patrones) repetibles del diseño arquitectónico que se encuentran en tipos similares de aplicaciones.
Dinámicos	Tratan los aspectos de comportamiento de la arquitectura del programa, indicando cómo puede cambiar la estructura o la configuración del sistema en función de los acontecimientos externos.
Proceso	Se centran en el diseño del proceso técnico de negocios que tiene que adaptar el sistema.
Funcionales	Se pueden utilizar para representar la jerarquía funcional de un sistema

Fuente: Tomado de Gómez (2011)

Crawlers

Según Rojas y otros (2009), un Crawler es un programa que permite recolectar páginas por cada sitios de la Web. Para ello, generalmente se usa un listado de sitios desde los cuales el proceso de recolección comienza.

Valera (2011) comenta que existen dos tipos de sistema de Crawler, dependiendo del propósito de la búsqueda, aquellos de propósito general y los de propósito especifico. Los de propósito general según Valera (2011) se incluye aquellos que recopilan documentos sin considerar algún tópico de búsqueda, además tienen como objetivo principal localizar y registrar el mayor número posible de páginas en La Web. Los de propósito específico, según Álvarez (2007) (citado por Valera 2011) son aquellos que localiza y obtiene páginas relacionadas con un

conjunto de temáticas determinadas, que representan segmentos relativamente limitados.

Rojas y otros (2009), asegura que dentro del grupo de crawler que son dirigidos se encuentran los que realizan el recorrido del grafo, el camino que ejecuta el algoritmo puede por profundidad o por amplitud. Para el caso de las herramientas de monitoreo de redes sociales Canali y otros (2011), asegura que existe una buena literatura acerca de Crawler dirigidos basados en el algoritmo breadth-first (amplitud - primero) o también llamado Breadth-Search-First (BSF) (Búsqueda en anchura), el cual los autores confirman que este es el tipo de crawler indicado para la recolección de datos en redes sociales.

Contenido Generado por Usuario (UGC)

Según la consultora Rooter (2011) define Contenido Generado por Usuario o su término en ingles *User-Generated Content* (UGC), aquellos contenidos creados por un usuario no profesional que no tienen fines comerciales directos o indirectos y que son divulgados, puestos a disposición del público o publicados a través de redes digitales.

Las redes sociales han sido los soportes que más han propiciado este intercambio de contenidos generados entre los usuarios.

Existen ciertas condiciones que deben cumplirse antes de catalogar un contenido en la web de tipo UGC. Primero el contenido tiene que ser contenido publicado, es decir, publicado por el usuario. Segundo tiene que ser de creación propia e individual y tercero no puede tratarse de contenido copiado ni reproducido.

Martínez (2007) en su artículo, resalta que existen particularidades más determinantes que permiten identificar a los UGC, estas se mencionan de manera detallada en el cuadro 4:

Cuadro 4
Característica de los Contenidos Generados por el Usuario*(*UGC)

Característica	Descripción
Propiedad de los Contenidos	Los *UGC* tienen claramente definido su propietario, pero en los *UGC* la propiedad no es clara entre el autor y el dueño de los medios que se utilizan en su difusión.
Calidad y Supervisión	Debido a que cualquier usuario puede crear *UGC* por múltiples vías es más difícil controlar la calidad, relevancia y veracidad del contenido publicado. La administración o moderación se convierte en un aspecto decisivo para supervisar la credibilidad de la información suministrada.
Estructura	Dependiendo del medio utilizado puede ser difícil estructurar, clasificar e indexar los *UGC*, haciendo muy difícil la búsqueda organizada, el control de la audiencia y la publicidad dirigida.

Fuente: Martínez (2007)

Repositorio de Datos

Según el sitio www.wikipedia.com, un repositorio, depósito o archivo es un sitio centralizado donde se almacena y mantiene información digital, habitualmente bases de datos o archivos informáticos. Según el sitio blog Poliscience, asegura que res tipos principales de repositorios:

- ✓ Repositorios institucionales: son los creados por las propias organizaciones para depositar, usar y preservar la producción científica y académica que generan. Supone un compromiso de la institución con

el acceso abierto al considerar el conocimiento generado por la institución como un bien que debe estar disponible para toda la sociedad.

- ✓ Repositorios temáticos: son los creados por un grupo de investigadores, una institución, etc. que reúnen documentos relacionados con un área temática específica.
- ✓ Repositorios de datos: repositorios que almacenan, conservan y comparten los datos.

Big Data

Según García (2013), big data son activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones.

García (2013), afirma que las característica que representa big data viene denominado por las tres 3V (velocidad, volumen, variedad), esto implica la complejidad de tratar estos tipos de datos. Esta complejidad viene derivada de la multiplicidad de fuentes de información que inciden sobre cualquier organización y están caracterizadas por estas 3 V. Según Fragoso (2012) existe una amplia variedad de tipos big data a analizar, una buena clasificación es la que se puede ver en la figura 8, en donde se ve mejor su representación, aunque Fragoso (2012) asegura que es muy probable que estas categorías puedan extenderse mas adelante con el avance tecnológico.

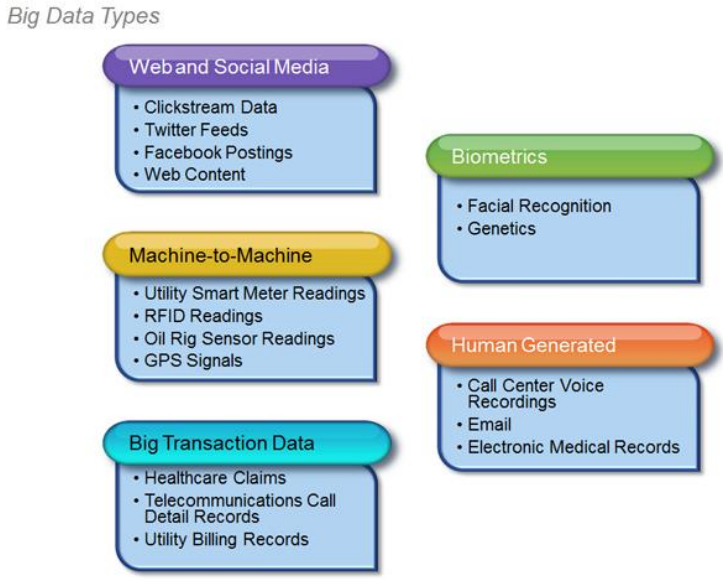


Figura 8 Tipos de Big Data
Fuente: Fragoso (2012)

A continuación se explica la clasificación mencionada por Fragoso (2012), detallada en la figura 8*:*

1.- Web and Social Media (Páginas web y social media): Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc, blogs.

2.- Machine-to-Machine (M2M) (Maquina a Maquina): se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

3.- Big Transaction Data (Datos de transacciones grande): Incluye registros de facturación, en telecomunicaciones los Call Detail Record (CDR) (registros detallados de llamadas) y otros. Estos datos transaccionales están disponibles en formatos tanto semi-estructurados como no estructurados.

4.- Biometrics (Biométricos): Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, entre otros. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.

5.- Human Generated (Generación Humana): Las personas generamos diversas cantidades de datos como la información que guarda un centro de llamadas al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, entre otros.

Minería de Datos

Según el sitio msdn.microsoft.com, define el data mining (minería de datos), como el proceso de detectar la información procesable de los conjuntos grandes de datos. Dicho proceso se apoya en análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos.

Molina (2002), asegura que el proceso se compone de cuatro etapas principales:

1. Determinación de los objetivos. Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en minería de datos.

2. Pre-procesamiento de los datos. Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa

consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de data mining.

3. Determinación del modelo. Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.

4. Análisis de los resultados. Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

Procesos ETL

Martínez (2013) los procesos ETL (Extract-Transform-Load) tiene como objetivo extraer la información de la fuentes, transformarla y cargarla en el almacén de datos.

La idea es que una aplicación ETL lea los datos primarios de unas bases de datos de sistemas principales, realice transformación, validación, el proceso cualitativo, filtración y al final escriba datos en el almacén y en este momento los datos son disponibles para analizar por los usuarios.

Martínez (2013) afirma que estos procesos se pueden dividir en tres subprocesos que se detallan a continuación:

- Extracción: consiste en extraer los datos desde los sistemas de origen. La mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Cada sistema separado puede usar una organización diferente de los datos o formatos distintos.

- **Transformación:** se aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Algunas fuentes de datos requerirán alguna pequeña manipulación de los datos.

- **Carga:** es el momento en el cual los datos de la fase anterior (transformación) son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. En algunas bases de datos se sobrescribe la información antigua con nuevos datos.

NoSql

Al hablar de NoSQL no se refiere de un tipo de base de datos sino de diferentes soluciones dadas para almacenar datos cuando las bases de datos relacionales no pueden satisfacer las especificaciones técnicas que se necesitan para llevar este tipo de operaciones. Según Mercado (2013), las bases de datos NoSQL son sistemas de almacenamiento de información que no cumplen con el esquema entidad-relación, mientras que las tradicionales bases de datos relacionales basan su funcionamiento en tablas, funciones CRUD, las bases de datos NoSQL no imponen una estructura de datos en forma de tablas y relaciones entre ellas (no imponen un esquema pre-fijado de tablas), lo cual en ese sentido son más flexible ya que suelen permitir almacenar información en otros formatos como clave-valor (similar a tablas Hash), Mapeo de Columnas, Documentos o Grafos.

Mercado (2013) continúa que la principal característica de las bases de datos NoSQL es que están pensadas para manipular enormes cantidades de información de manera muy rápida, para ello suelen almacenar toda la información que pueden en memoria (utilizando el disco como una mera herramienta de persistencia), y están preparadas para escalar horizontalmente sin perder rendimiento.

Continuando con lo anterior Díaz (2013), comenta que las bases de datos NoSQL intentan resolver problemas de almacenamiento masivo, alto desempeño, procesamiento masivo de transacciones (sitios con alto transito) y en términos generales ser alternativas NoSQL a problemas de persistencia y almacenamiento masivo (voluminoso) de información para las organizaciones.

Cuervo y Sanabria (2012) destaca las principales características que posee un almacenamiento NoSQL tenemos:

- **Consistencia Eventual:** No se implementan mecanismos rígidos de consistencia como los presentes en las bases de datos relacionales, donde la confirmación de un cambio implica una comunicación del mismo a todos los nodos que lo repliquen. Esta flexibilidad hace que la consistencia se dé, eventualmente, cuando no se hayan modificado los datos durante un periodo de tiempo.

- **Estructura distribuida**: Generalmente se distribuyen los datos mediante mecanismos de tablas de hash distribuidas.

- **Libertad de esquema**: al no tener un esquema rígido se permite mayor libertad para modelar los datos; además facilita la integración con los lenguajes de programación orientados a objetos, lo que evita el proceso de mapeado, todo esto origina que el esquema de los datos manejado aquí posee un estructura dinámica.

- **Escalabilidad horizontal:** Consiste en la posibilidad de aumentar el rendimiento del sistema simplemente añadiendo más nodos, sin necesidad en muchos casos de realizar ninguna otra operación más que indicar al sistema cuáles son los nodos disponibles. Muchos sistemas NoSQL permiten utilizar consultas del tipo Map-Reduce, las cuales pueden ejecutarse en todos los nodos a la vez y reunir luego los resultados antes de devolverlos.

- **Modelo concurrencia débil:** no implementa atomicidad, coherencia, aislamiento y durabilidad (ACID), que reúne las características necesarias para que una serie de instrucciones puedan ser consideradas una transacción, sin embargo sí se tienen en cuenta algunas consideraciones para asegurar estos aspectos, pero no son tan estrictas.

- **Consultas simples:** las consultas requieren menos operaciones y son más naturales, por la tanto, se gana en simplicidad y eficiencia. Esto origina así evitan la generación de cuellos de botella.

Dentro del grupo de soluciones que abarca NoSql se aplican tecnologías como Hadoop, Distributed File System (HDFS) y Map Reduce, lo que permite trabajar con ficheros de gran tamaño y proporciona un sistema de procesamiento de datos paralelo y distribuido. Según Díaz (2013) en su artículo resalta que este tipo de base de datos se han venido clasificando principalmente en cuatro (4) grupos: De Clave Valor, Documentos, Familia de columnas y Grafos. En el cuadro 5 se describen brevemente los tipos de base de datos NoSql .

Cuadro 5
Clasificación de Base de Datos NoSQL

Clasificación	**Característica**	**Ejemplos**
De Clave Valor	Este grupo de bases de datos NOSQL cuyo precursor fue Big Table de Google tiene un Modelo con pares clave-Valor Especialmente útiles para problemas de escrituras masivas de "Streaming". Transacciones tipo son : put (key, value), get(key), remove(key)	Dynamo Amazon, Cassaandra, Voldemort, Redis.
Documentos	Las bases de datos de este grupo permiten la gestión de información semi-estructurada orientadas a documentos, son similares a registros, direccionados por una clave única, y se pueden recuperar con su contenido. Tienen un modelado muy natural	Couchdb, Mongodb,

	orientado a la web.	riak
Familia de columnas	Son almacenamientos de datos orientados a Columnas.	Casandra, Hbase
Grafos	Los nodos son entidades y los arcos con relaciones y contienen información con uso a menudo de tablas hash distribuidas y ofrecen estructuras de datos sencillas como arrays asociativos o almacenes de pares claves valor.	Neo4j, Flockdb

Fuente: Díaz (2013)

Teorema CAP

Según García (2013) en el mundo de los sistemas NoSQL se suele aplicar el Teorema CAP a la hora de elegir el tipo de base de datos a trabajar. Según este teorema existen 3 conceptos primarios que se debe estar alerto cuando se elija un una base de datos, estas propiedades (ver figura 9) son:

-Consistencia: representa la característica en que cada cliente siempre tiene la misma vista de los datos.

-Alta Disponibilidad: indica que todos los clientes puedan siempre leer y escribir.

-Tolerancia a particionamiento: representa que el sistema funciona bien entre particiones de redes físicas.

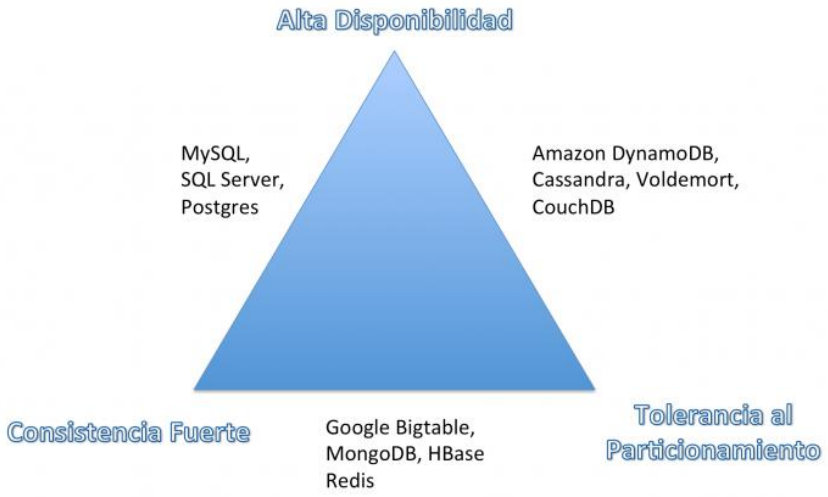


Figura 9. Teorema CAP para sistema distribuido
Fuente: Martin (2013)

Modelado en NoSQL

Katsov (2012) sugiere técnicas y patrones de modelado para el almacenamiento de datos no estructurados, a continuación primero se presenta los principios básicos de modelado de datos NoSQL y luego se presentan las técnicas de generales de modelado propuesta por Katsov (2012).

Principios Básicos de Modelado de Datos NoSQL

Katsov (2012) menciona, que existen tres principios básicos, que se observa en un modelo basado en NoSQL:

1) Desnormalización

La desnormalización la define Katsov (2012) como la copia de los mismos datos en varios documentos o tablas con el fin de simplificar o optimizar el procesamiento de consultas, con el fin de ajustar los datos del usuario en un modelo de datos en particular. El autor señala que la desnormalización es útil para las siguientes concesiones:

- Volumen de entradas o salidas por consulta versus el volumen total de datos : Al aplicar la técnica de desnormalización, se puede agrupar todos los datos necesarios para procesar una consulta en un solo lugar. Esto a menudo significa que los flujos diferentes de consultas de los mismos datos se accederá en diferentes combinaciones. Por lo tanto se debe duplicar los datos, que aumenta el volumen total de datos.

- Complejidad de procesamiento versus volumen total de datos: La normalización en tiempo de modelado y el consiguiente tiempo de consulta, aumentan la complejidad del procesador de consultas, sobre todo en los sistemas distribuidos. La desnormalización permite almacenar los datos en una estructura de consulta amigable para simplificar el procesamiento de consultas.

Katsov (2012) comenta que esta técnica es aplicable en almacenamientos de de tipo valor-clave, documentos y big table (tablas grandes).

2) Los Agregados

Como lo define Fowler y Pramod (2012) un agregado es una colección de objetos relacionados que se desea tratar como una unidad, con la finalidad de gestionar y manipular la unidad para mantener la consistencia en los datos. Los autores coincide que esta definición se adapta muy bien a los almacenamientos de clave-valoy y a las bases de datos orientadas a documentos y columnas, ya que esta técnica permite que sea muy facil trabajar los clúster operativo que se encuentran

dentro de las bases de datos, ya que el agregado como unidad natural sirve de replicación y compartimiento

Fowler y Pramod (2012) afirma que los agregados también son a menudo más fáciles para los programadores a la hora de trabajar con aplicaciones, ya que menudo ellos manipulan los datos a través de estructuras globales.

Por otro lado Katsov (2012) comenta que todos los modelos de NoSQL proporcionan capacidades de esquema suaves en una u otra forma, lo que permite formar clases de entidades con complejas estructuras internas (entidades anidadas) y variar la estructura de entidades particulares. Esta característica proporciona dos comodidades principales:

- Reducción al mínimo de relaciones de uno-a-muchos, por medio de entidades anidadas lo cual genera la reducción de relaciones de tipo unión.

- Enmascaramiento de las diferencias técnicas entre las entidades empresariales y modelos de entidades heterogéneos utilizando una colección de documentos o una tabla.

Estas comodidades se ilustran en figura 10, en donde representa el modelado de una entidad denominada "producto" de un dominio de negocio de "comercio electrónico". Inicialmente, se tiene que todos los productos tienen los siguientes atributos: identificación, precio y descripción. A continuación, se descubre que los diferentes "tipos de productos" tienen otros atributos específicos como: autor de libro o longitud de los pantalones, algunos de estos atributos tienen una relación de tipo uno-a-muchos o muchos-a-muchos. Inmediatamente, se observa que algunas entidades no se pueden modelar usando atributos fijos para todo. Por ejemplo, los atributos de los pantalones vaqueros no son consistentes a través de las marcas y tampoco son específicos para cada fabricante. Como lo comenta Katsov (2012) es posible superar todos estos obstáculos en un modelo de datos relacional normalizado, pero la solución podría estar lejos de ser elegante. El esquema suave resaltado por

Katsov (2012) permite utilizar un único agregado en este caso "productos" que pueden modelar todo tipo de productos y sus atributos, tal como se muestra en la figura 10.

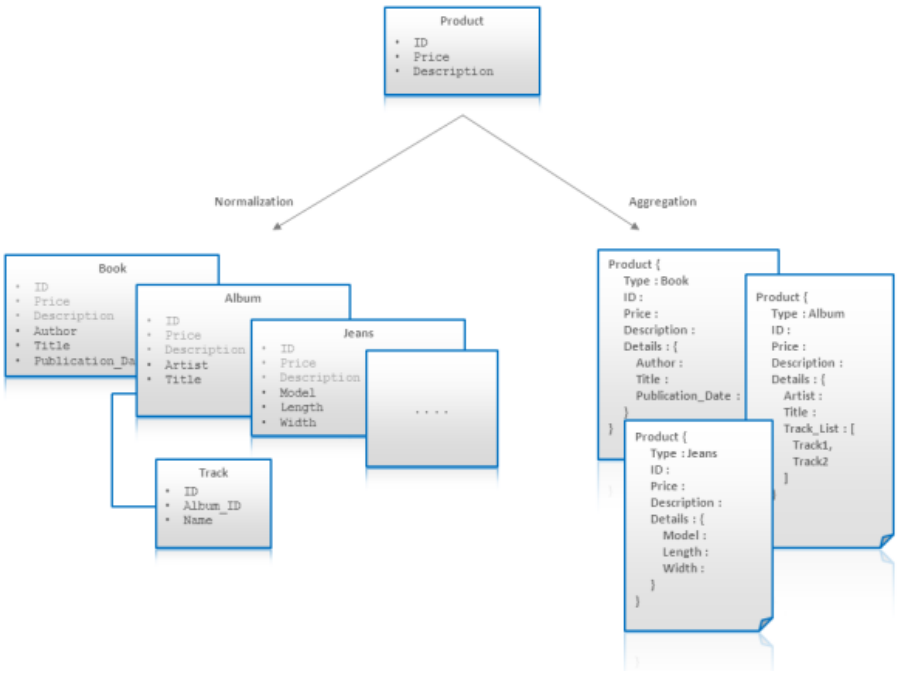


Figura 10. Entidad de Agregación**.**
Fuente: Katsov (2012)

Katsov (2012) comenta que esta técnica es aplicable en almacenamientos de de tipo valor-clave, documentos y tablas grandes.

3) Aplicación de Uniones Laterales

Las combinaciones (joins) son soportada rara veces en soluciones NoSQL. Katsov (2012) asegura que la naturaleza de NoSQL es orientado a preguntas, por lo tanto las combinaciones en NoSQL son manejadas a menudo en tiempo de diseño, a diferencia de los modelos relacionales, donde las combinaciones se hace en tiempo

de ejecución de la consulta. Katsov (2012) comenta que las combinaciones en tiempo de consulta casi siempre significa una penalización en el rendimiento, pero en muchos casos se puede evitar el uso de combinaciones, desnormalización y agregados, es decir, la incorporación de entidades anidadas. Por supuesto, en muchos las combinaciones son inevitables y deben ser manejados por una aplicación. Los casos de mayor uso son:

- En relaciones de muchos a muchos donde a menudo se modela mediante enlaces los cuales requieren uniones.

- Los agregados son a menudo inaplicable cuando entidades internas son objeto de frecuentes modificaciones. Por lo general, es mejor mantener un registro de que algo pasó y unir los registros en tiempo de consulta en lugar de cambiar un valor. El ejemplo mostrado por Katsov (2012), muestra un sistema de mensajería que puede ser modelado como una entidad de usuario que contiene entidades de mensajes anidados, tomando en cuenta si los mensajes se anexan a menudo, puede ser mejor extraer los mensajes como entidades independientes y unirlas a la entidad usuario en tiempo de consulta, dicha descripción se observa en la figura 11.

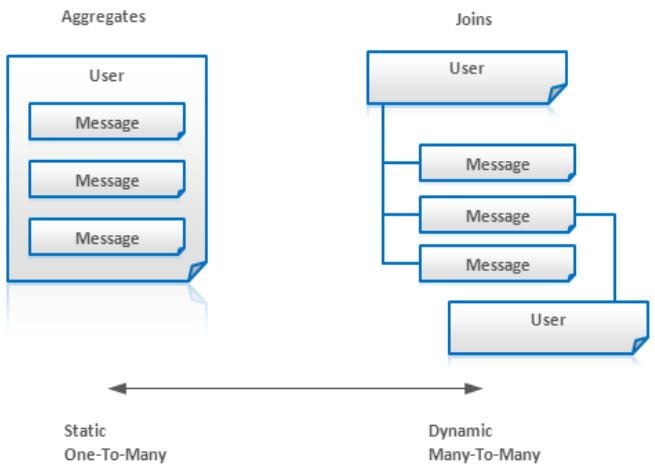


Figura 11. Ejemplo de utilización de entidades anidadas**.**
Fuente: Katsov (2012)

Técnicas de Generales de modelado NoSQL

Entre las técnicas que menciona Katsov (2012), están:

1) Los agregados atómicos

Katsov (2012), resalta que muchas soluciones NoSQL aunque no todas, tienen limitado el soporte de transacciones. En algunos casos se puede almacenar un comportamiento transaccional mediante bloqueos distribuidos o aplicación gestionando MVCC (Control de concurrencia multiversión), pero es común para modelar los datos la utilización de agregados para garantizar algunas de las propiedades ACID.(Atomicidad, Consiatencia, Aislamiento y Durabilidad).

Una de las razones asegura Katsov (2012), por las cuales la poderosa maquinaria transaccional es una parte inevitable de las bases de datos relacionales es que los datos normalizados suelen requerir múltiples actualizaciones en el lugar. Por otra parte, los agregados permiten almacenar una sola entidad de negocios en un solo documento, en conjunto con su clave-valor para actualizarlo de forma atómica, dicha descripción se puede ver en la figura 12.

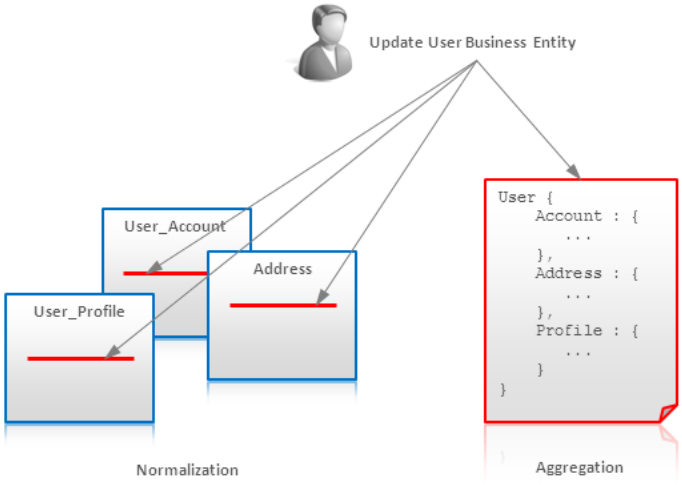


Figura 12. Ejemplo de agregados atómicos.
Fuente: Katsov (2012)

Katsov (2012), asegura que los agregados atómicos como una técnica de modelado de datos no es una solución transaccional completa, pero si el almacenamiento ofrece ciertas garantías de atomicidad, bloqueos, o instrucciones de TSL (Test and set lock) (prueba y sistema de bloqueo) entonces los agregados atómicos pueden ser aplicables.

2) Claves enumeradas

Katsov (2012), comenta que el mayor beneficio de un modelo de datos desordenado de clave y valor es que las entradas se pueden repartir entre varios servidores numerando solo la llave. La clasificación hace las cosas más complejas, pero algunas veces las aplicaciones toman ventaja de llevar llaves ordenadas, aunque el almacenamiento no ofrezca tal característica. Katsov (2012) considero como ejemplo el modelado de mensajes de correo, a continuación se detalla por pasos el ejemplo mencionado:

1) Algunos almacenamientos NoSQL proporcionan contadores atómicos que permiten generar IDs secuenciales. Katsov (2012) asegura en este caso que se puede almacenar mensajes utilizando userIDmessageID como una clave compuesta. Si se conoce el último ID de mensaje, es posible recorrer los mensajes previos. También es posible recorrer los mensajes anteriores y posteriores para cualquier ID de mensaje dado.

2) Los mensajes se pueden agrupar internamente en cubos, por ejemplo, cubos diarios. Esto permite atravesar una casilla de correo hacia atrás o hacia delante a partir de cualquier fecha específica o la fecha actual .

Katsov (2012) comenta que esta técnica solo es aplicable en almacenamientos de tipo valor-clave.

3) Reducción de Dimensionalidad

Reducción de dimensionalidad es una técnica que permite cartografiar datos multidimensionales a un modelo de valor-clave o para otros modelos que no sean multidimensionales, un ejemplo de este tipo de modelo son los sistemas de información geográfica .

Los sistemas de información geográfica tradicionales, utilizan alguna variación de un árbol cuádruple (Quadtree) o R-Tree (R-árbol) para los índices. Estas estructuras necesitan ser actualizadas en el lugar y son caros para manipular cuando los volúmenes de datos son grandes. Un enfoque alternativo es atravesar la estructura 2D y aplanar en una lista simple de entradas. Un ejemplo bien conocido de esta técnica es un Geohash. Un Geohash utiliza una exploración Z como para llenar el espacio 2D y cada movimiento se codifica como 0 o 1 dependiendo de la dirección. Bits de longitud y latitud se mueve se intercalan y se mueve. El proceso de codificación se ilustra en la siguiente figura 13.

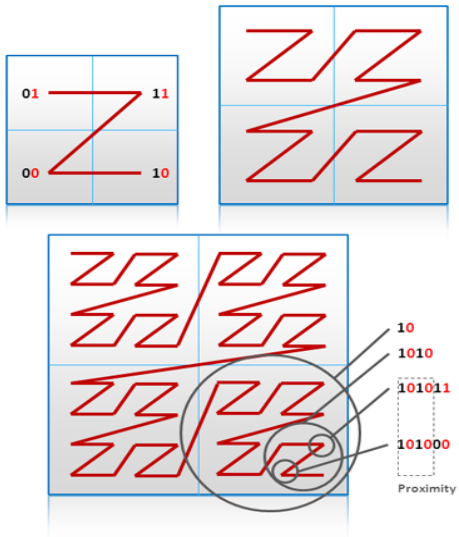


Figura 13. Ejemplo de Geohash Índice.
Fuente: Katsov (2012)

Katsov (2012) comenta que esta técnica es aplicable en almacenamientos de de tipo valor-clave, documentos y tablas grandes.

4) Tabla Índice

Katsov (2012) la define como una técnica muy sencilla que permite sacar ventaja de los índices en almacenamientos que no soportan índices internamente. Katsov (2012) asegura que la idea es crear y mantener una tabla especial con claves que siguen el patrón de acceso. Katsov (2012) indica un ejemplo en donde describe una tabla maestra que almacena las cuentas de usuario que se puede acceder por el ID de usuario, seguidamente la consulta que recupera todos los usuarios de una determinada ciudad se apoya a través de una tabla adicional donde la ciudad es la clave, la figura 14 muestra el ejemplo anteriormente descrito.

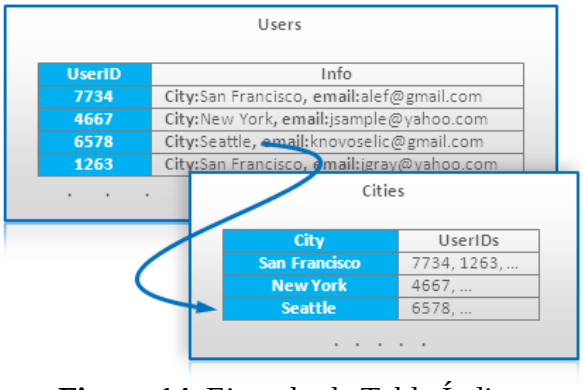


Figura 14. Ejemplo de Tabla Índice.
Fuente: Katsov (2012)

Katsov (2012) asegura que una tabla de índice se puede actualizar para cada actualización de la tabla principal o en modo lote. De cualquier manera, el resultado es una reducción del rendimiento adicional y convertido en una cuestión de coherencia. Katsov (2012) comenta que la tabla de índice puede ser considerado como un análogo de vistas materializadas en las bases de datos relacionales.

Katsov (2012) comenta que esta técnica solo es aplicable en almacenamientos orientados a tablas grandes.

5) Índice clave Compuesta

Katsov (2012) comenta que esta técnica es muy genérica, pero es extremadamente beneficioso cuando se utiliza en almacenamientos con claves ordenadas. Katsov (2012) continua diciendo que las claves compuestas en combinación con la clasificación secundaria permite al usuario construir una especie de índice multidimensional que es fundamentalmente similar a la técnica de reducción de dimensionalidad . Katsov (2012) describe un ejemplo, donde se obtiene un conjunto de registros, donde cada registro es una estadística de usuario, luego si agregas estas estadísticas por región donde se indica la procedencia del usuario, se puede utilizar las claves en un formato como "Estado: Ciudad: Usuario", esto nos va a permitir recorrer los registros para un estado o ciudad en particular si ese almacenamiento es compatible con la selección de rangos de clave por una coincidencia parcial de la clave seleccionada, en la figura 15 muestra el ejemplo descrito.

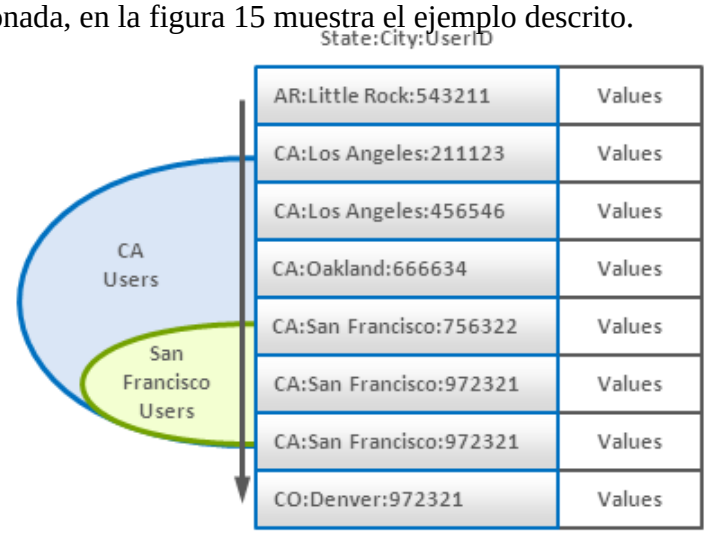


Figura 15. Ejemplo de Indice clave Compuesta.
Fuente: Katsov (2012)

Plataforma de Agregación

Las plataformas de agregación son sitios en donde utilizan técnicas de agregación de datos, para mostrar la información de manera ágil y unificada. Rouse (2005) define la agregación de datos como el proceso en el que la información se recopila y se expresa en una forma resumida, para fines tales como el análisis estadístico.

Rouse (2005) confirma que el propósito común de la agregación de datos es obtener más información acerca de grupos particulares basados en variables específicas tales como la edad, la profesión, o ingresos. La información obtenida sobre estos grupos entonces puede utilizarse para la personalización del sitio web para mostrar publicidad probable para atraer a un individuo perteneciente a uno o más grupos para los cuales se han recogido datos y contenidos. Por ejemplo, un sitio que vende discos compactos de música podría anunciar algunos CDs basados en la edad del usuario y los datos agregados para su grupo de edad.

En el caso de la información que se generan en redes sociales, los sitios que funciona como herramientas de monitoreo de redes sociales basada en técnicas de agregación, según el sitio www.wikipedia.com, lo define como el proceso de recoger el contenido de varios servicios de redes sociales, como MySpace o Facebook, en una presentación unificada. La tarea se realiza a menudo por una red social que funciona como recolector, que reúne la información en un solo lugar y ayuda a los usuarios a consolidar los múltiples perfiles de redes sociales en un solo perfil.

Componentes de Software

Los componentes de software como lo definen Rojas y García (2004) es una unidad de composición con interfaces contractualmente especificadas y explícitas sólo con dependencias dentro de un contexto. Los autores continua que existen

algunas características claves para que un elemento pueda ser catalogado como componente, tales como: identificable, auto contenido, reemplazable, con acceso solamente a través de su interfaz, sus servicios no varían, documentado, genérico y finalmente independiente de la plataforma.

Agentes

Un agente es una abstracción que describe una pieza de software que actúa sobre un tercero (usuario, programa, entre otros) dentro de una relación con otros agentes. Para Hyacinth, Nwana y Ndumu (2006) existen diferentes definiciones para los agentes de software, sino que abarcan la mayoría de ellos las siguientes características:

• Autónomo - Los agentes pueden seleccionar las tareas a realizar, puede tomar decisiones sin intervención externa.

• Reactivo - Los agentes perciben el contexto en el que operan y reaccionan a de manera apropiada.

• Proactivo - Los agentes no lleva a cabo en orden y decide por sí mismo, cuando debería realizar alguna actividad**.**

• Meta - Impulsado por los agentes se comportan sobre la base de un objetivo con que se define explícitamente, este es el desencadenante de su interacción.

• Social – Los agentes son capaces de participar comunicación con el agente medio ambiente para colaborar en una tarea.

• Adaptable – Los agentes están analizando el entorno y reaccionar en consecuencia.

• Cognitivo – El aprendizaje puede ocurrir a través de ensayo y error o la generalización.

Abarca la capacidad de un agente para reaccionar sobre la base de las experiencias pasadas.

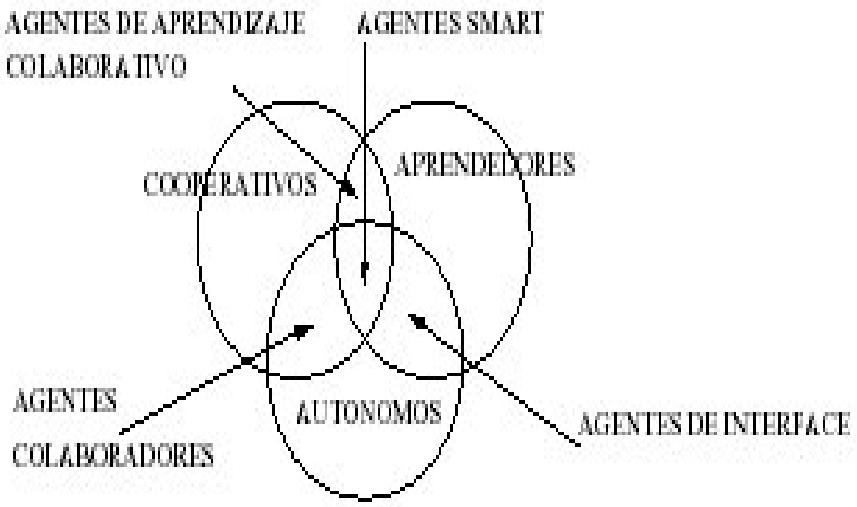


Figura 16. Clasificación de Agentes según Hyacinth S. Nwana, C, Ndumu, T.
Fuente: Hyacinth y Nwana (2006)

CAPITULO III

MARCO METODOLOGICO

A lo largo de este capítulo se pretende especificar la metodología de investigación que será empleada para el desarrollo de este trabajo, cuyo fin es responder satisfactoriamente los objetivos planteados en el problema. En el mismo se desarrollaran los aspectos metodológicos en los cuales estará enmarcada la investigación, en la cual es necesario realizar un análisis de los temas estudiados con el fin de generar respuestas y conclusiones que aporten algún valor para solventar los inconvenientes y detalles por resolver en los procesos de obtención y administración de información proveniente de las diferentes redes sociales y presentar una solución que logre el almacenamiento de estos contenidos sociales generados, para garantizar procesos de análisis y monitoreo.

Naturaleza de la Investigación

Por la naturaleza de esta investigación, la cual propone definir una arquitectura que garantice la construcción de un almacenamiento de datos generados en redes sociales para proceso de análisis y monitoreo, se presenta bajo la modalidad de proyectos tal como lo especifica el manual UCLA (2002). Para ello, el trabajo se

enmarca dentro de la modalidad de Proyecto Especial; porque la definición de la propuesta coadyuvará en mejorar estrategias de indexación que faciliten la exploración de la información generada en redes sociales, fomentando así que docentes, estudiantes, investigadores y usuarios en general puedan satisfacer sus necesidades de búsqueda de información generada en su entorno social.

Al respecto de esta modalidad, la Universidad Centroccidental "Lisandro Alvarado" (UCLA) (2002), en su Manual para la Elaboración del Trabajo Conducente a Grado Académico de Especialización, Maestrías y Doctorado, la presente investigación se ubica en la modalidad de Estudios de Proyectos, el cual consiste en "una proposición sustentada en un modelo viable para resolver un problema práctico planteado, tendente a satisfacer necesidades institucionales o sociales y pueden referirse a la formulación de políticas, programas, tecnología, métodos y procesos" (p. 63).

Del mismo modo, el Manual de Trabajos de Grado de Maestría y Tesis Doctorales de la Universidad Pedagógica Experimental Libertador (UPEL) (2006), define el proyecto especial como: "Trabajos que lleven a creaciones tangibles, susceptibles de ser utilizados como soluciones a problemas demostrados, o que respondan a necesidades e intereses de tipo cultural". (p. 14).

De esta manera, todo proyecto especial debe incluir la demostración de la necesidad de la creación o de la importancia del aporte; por lo que la presente metodología se apoya, en su fase de análisis , en una investigación documental. Así como lo afirma Cázares y otros (1980), "La investigación documental depende fundamentalmente de la información que se recoge o consulta en documentos, entendiéndose este término, en sentido amplio, como todo material de índole permanente, es decir, al que se puede acudir como fuente o referencia en cualquier momento o lugar, sin que se altere su naturaleza o sentido, para que aporte información o rinda cuentas de una realidad o acontecimiento", (p. 18).

El Manual de la UPEL, establece que una investigación documental, de acuerdo a los objetivos del estudio o de la temática, pueden ser:

> ...Revisiones críticas del estado del conocimiento: integración, organización y evaluación de la información teórica y empírica existente sobre un problema, focalizando ya sea en el progreso de la investigación actual y posibles vías para su solución, en el análisis de la consistencia interna y externa de las teorías y conceptualizaciones para señalar sus fallas o demostrar la superioridad de unas sobre otras, o en ambos aspectos... (p. 13).

En este sentido, se evaluarán los distintos métodos y herramientas que existen actualmente para la implementación de soluciones de captura y almacenamiento de datos en redes sociales, así como también los desafíos que subsisten en este entorno para gestionar datos no estructurados, para comparar así dicha tecnologías y métodos con la finalidad de proponer una solución que se apoye en estándares reconocidos.

Método de Recolección de Información

En la presente investigación, es imprescindible el uso de herramientas o instrumentos que permitan una recolección de datos efectiva y eficaz, que garantice una base sólida para el análisis y el éxito del estudio. Es por tal motivo, se aplicaron técnicas de recopilación documental, semántica documental y análisis de contenido; haciendo uso de recursos de documentos escritos, como libros, revistas y tratados; documentos electrónicos como páginas web, revistas digitales, presentaciones y conferencias. Cuyo objetivo principal es el acopio de los antecedentes relacionados con la investigación, en donde para tal fin se consultaron documentos escritos formales que se tomaron como base y enfoque para la presente investigación; incluso se admitieron ideas, mecanismos, propuestas ya probados en investigaciones y/o proyectos anteriores.

Diseño de la Investigación

Fases de la Investigación

A fin de cumplir con los requisitos involucrados en la modalidad, la cual se establece como Proyecto Especial y en función de los objetivos planteados en la investigación, se determinaron 3 fases para el estudio:

Fase I: Fase de Diagnóstico y Análisis.

En esta fase se desarrolla el conocimiento de la situación existente en la realidad objeto de estudio, a fin de describir las funcionalidades requeridas por la arquitectura de software a proponer, apoyado en material bibliográfico a través de un proceso sistemático de recolección, organización, análisis e interpretación de información

El propósito de esta fase es el de analizar y diagnosticar la problemática existente, referente al manejo de los datos generados en redes sociales, identificando así los métodos actuales de recuperación que se emplean en los distintos contenidos que se generan en los espacios sociales. Cabe destacar también que es necesario identificar las características del repositorio de datos no estructurados que permita almacenar datos de redes sociales para realizar procesos posteriores de análisis y monitoreo. Entre las tareas que se ejecutaron en esta fase se encuentran:

- Búsqueda Bibliográfica: obtención de información en base a los trabajos desarrollados referente a modelos arquitecturales que manejan datos de redes sociales, como también fuentes que refieran técnicas o tecnologías para definir repositorio de datos no estructurado .

- Énfasis de los temas fundamentales en donde la presente investigación tiene su mayor peso**.**

- Identificación de los métodos de recuperación que se emplean para gestionar datos de redes sociales, esto tomado a partir de los trabajos relacionados en donde se

especifican las técnicas utilizadas para obtener los datos generados en redes sociales. Adicional se determinaron los procesos que están relacionados con la exploración y recolección de estos tipos de datos.

Especificación de Requisitos: Este proceso se realizó tomando el modelo de requisitos presentado por Boukhanovsky y Souravlias (2012), en el cual refleja una caracterización de una solución que gestiona datos de diferentes redes sociales, adicional se tomó el modelo presentado por Mysore y otros (2013) como guía, en donde se identifican los elementos asociados al proceso de construcción, alimentación y extracción de un repositorio de datos en el marco de las redes sociales

Fase II: Determinar las características del repositorio de datos no estructurado donde almacenará los datos obtenido de redes sociales

En base al análisis realizado en la fase anterior, se describen las principales características que debe presentar el repositorio, para almacenar datos no estructurados generados en redes sociales, especificando así, las tecnologías seleccionadas para garantizar el acceso y recuperación de la información almacenada, igualmente se describe la estructura que debe presentar los datos, para que puedan ser almacenados.

Fase III: Fase de Diseño de Arquitectura

En esta tercera fase, tendrá como insumo el resultado de la fase I y II, ya que se obtiene una información amplia y concisa de las características, recursos disponibles, descripción y requisitos de la arquitectura. En esta fase se determinó los componentes que interactuaran en la arquitectura siguiendo los trabajos de Boukhanovsky y Souravlias(2013) , Mysore y otros (2013) , dicha arquitectura estará estructurada por dos capas : la primera se encargará de extraer los datos de las diferentes fuentes sociales implementando uno o varios mecanismos de recuperación indicado por Canali (2011), los cuales interactuaran con las API publicas de cada red social, la

segunda capa se encargara de gestionar el almacenamiento de los datos previamente obtenidos.

CAPITULO IV

ANALISIS DE LOS RESULTADOS

Este capítulo contempla el diseño de la arquitectura para la conformación de un repositorio de datos provenientes de redes sociales para procesos de análisis y monitoreo, es importante señalar que en este capítulo se desarrollan los pasos que determinaron el desarrollo de el diseño de la arquitectura a través de las fases señaladas en el capitulo metodológico, demostrando así que el uso de estos datos pueden ayudar a conformar un almacenamiento de datos no estructurados para aplicar técnicas de análisis y monitoreo que ayuden a ejecutar procesos de preparación y exploración de datos, cubierta por la minería de datos, con el propósito de generar análisis exactos en base a la información que se genera en estos espacios sociales, los cuales ayuden a las organizaciones aumentar su conocimiento colectivo con el fin de mejorar su bases estratégicas para la toma decisiones.

Fase I. Diagnóstico y Análisis

En esta etapa se profundiza en cada uno de los aspectos que conforman las bases teóricas y se hará principal énfasis en seis (6) temas fundamentales en los cuales la presente investigación tiene su mayor peso, para así cumplir el propósito final de la

investigación, como es el diseño de una arquitectura de software para la conformación de un repositorio de datos de redes sociales para procesos de análisis y monitoreo. Estos temas son: big data, arquitectura big data, NoSQL, sistema multiagente y arquitectura orientada a servicios.

También se identificaron los mecanismos de búsqueda y recuperación que se usa para los diferentes contenidos que se genera en las redes sociales, los cuales se consideraron para el diseño de la arquitectura propuesta.

Redes Sociales

Castañeda y Gutierrez (2010) la definen como el conjunto de herramientas telemáticas de comunicación que tienen como base la web, que permiten a un usuario crear un perfil de datos sobre si mismo en la red y compartirlo con otros usuarios. Dicha herramientas tienen como objetivo conectar sucesivamente a los propietario de dichos perfiles a través de categorías, grupos, preferencias personales, entre otros ligado a su propia persona o perfil.

Castañeda y Gutierrez (2010) confirma que la base de estas herramientas es la red en si misma, donde se observa las características de las personas que están conectadas y los contenidos que generan en la red.

Codina (2009) asegura que el concepto de red social el concepto de red social precede a la Web. Se trata de una estructura social que puede adoptar muchas formas y diversas características, es decir se maneja en otros ámbitos de la ciencia como son la sociología, la documentación y la filosofía . En el cuadro 6 se presenta los componentes descrito por Codina (2009) que debe presentar una red social en el ámbito de web 2.0.

Cuadro 6

Componentes generales de la redes sociales en el contexto Web 2.0

Componente	Descripción
Página de inicio	Es la página que muestra la red cuando nos identificamos (login). Incluye las novedades que nos afectan (mensajes, visitantes, nuevas incorporaciones, etc.) y el acceso a las funciones básicas de la Red.
Perfil	La información que hemos decidido publicar sobre nosotros mismos. Puede ser muy básica o puede ser exhaustiva.
Mensajes	Un archivo de los mensajes anteriores con miembros de nuestra red personal
Búsquedas	Una función que permite buscar en la red por nombres de personas o por temas para encontrar nuevos contactos de la Red o grupos de interés
Grupos	Acceso a los grupos de los que formamos parte y opciones para buscar grupos, para solicitar formar parte y para crearlos. En el caso de redes académicas, obviamente los grupos son de este tipo.
Comunicación/Colaboración	Diversas herramientas que facilitan la comunicación y la colaboración entre grupos y miembros de la red
Preferidos	Algunas redes sociales permiten mantener una lista de sitios, páginas y recursos preferidos que otros miembros de la red también pueden consultar

Fuente: Codina (2009)

Codina (2009) resalta el aspecto más conocido de las redes sociales en el ámbito de la Web 2.0 , lo constituye no solamente el número de afiliados con los que

cuentan, sino el simple hecho de que son el único elemento que ha pasado a formar parte del universo de los medios de comunicación.

Big Data

Barranco (2012) continua detallando que las tres V, se debe a que el big data es una tecnología que posee gran volumen de información, de la cual existe una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, lo que implica que pueden venir de diversas fuentes, originando así que las aplicaciones que analizan estos datos requieran de una velocidad de respuesta rápida para lograr obtener la información correcta en el momento preciso.

La clasificación de los tipos de datos que se presenta en big data, están relacionado a la fuente de datos donde se generan, sin embargo como lo confirma Fragoso (2012) es muy probable que estas categorías puedan extenderse con el avance tecnológico. A continuación se presenta la clasificación presentada por Fragoso (2012), que describe los tipos de datos que se manejan en big data:

- Web and Social Media (Páginas Web y Social Media): Incluye contenido de páginas web y información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, blogs, entre otros.

- Machine-to-Machine (M2M) (Maquina a Maquina): M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

- Big Transaction Data (Datos de transacciones grande): Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (*Call Detail Record)* (CDR), entre otros. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

- Biométricos (Biometrics): Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, entre otros. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.

- Generación humana (Human Generated): Las personas generamos diversas cantidades de datos como la información que guarda un centro de llamadas al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, entre otros.

Arquitectura Big Data

En el ámbito de big data los expertos señalan que es posible almacenar, adquirir, procesar y analizar de muchas formas. Cada fuente de big data tiene distintas características, que incluyen la frecuencia, el volumen, la velocidad, el tipo y la veracidad de los datos. Cuando se procesan y almacenan los datos en big data, entran en juego dimensiones adicionales, como la gobernabilidad de los datos, la seguridad y las políticas. Elegir una arquitectura y desarrollar una solución basada en big data es un reto, ya que se deben considerar muchos factores**.**

Mysore y otros (2013) define las características básicas que debe presentar una solución big data, pero antes de identificarlas, los autores señalan que se debe realizar un estudio previo de ciertos factores como son: la recopilación, el análisis y el procesamiento de los datos. Mysore y otros (2013) aseguran que una vez se tenga definido los criterios anteriormente mencionados se procede a definir la arquitectura big data apropiada, en donde deben aparecer las siguientes propiedades:

- **Tipo de Análisis:** especificar si los datos se analizan en tiempo real o se organizan por lotes para ejecutar su análisis posterior.

- **Metodología de procesamiento:** El tipo de técnica que se aplicará para procesar los datos (por ejemplo, predictiva, analítica, consulta ad hoc e informes). Los requisitos empresariales determinan la metodología de procesamiento apropiada. Se puede utilizar una combinación de técnicas. La elección de la metodología de procesamiento ayuda a identificar las herramientas y técnicas apropiadas para utilizar en la solución de big data.

- **Frecuencia y tamaño de los datos**: Cuántos datos se esperan y con qué frecuencia llegan. Conocer la frecuencia y el tamaño ayuda a determinar el mecanismo y el formato de almacenamiento, y las herramientas de procesamiento necesarias. La frecuencia y el tamaño de los datos dependen de los orígenes de datos.

- **Formato del contenido:** Formato de los datos entrantes, los cuales pueden ser estructurados (RDMBS, por ejemplo), no estructurados (audio, video e imágenes, por ejemplo) y semiestructurados. El formato determina la forma en que los datos entrantes deben ser procesados y es clave para elegir las herramientas y técnicas, así como también para definir una solución desde una perspectiva empresarial.

- **Origen de datos**: Orígenes de datos representa las fuentes donde se generan los datos, por ejemplo: páginas web, redes sociales, máquinas, generados por humanos a través de las organizaciones, entre otros. Mysore y otros (2013) destaca que identificar todos los orígenes de datos ayuda a determinar el ámbito de una solución big data desde una perspectiva empresarial.

- Consumidores de datos: Una lista de todos los posibles consumidores de los datos que son gestionados, es decir quien utiliza los datos después que son procesado por la arquitectura big data:

- Procesos empresariales.
- Usuarios empresariales.
- Aplicaciones empresariales.
- Personas individuales en diversos roles empresariales.
- Parte de los flujos de proceso.
- Otros repositorios de datos o aplicaciones empresariales.

NoSQL

Cuando hablamos de NoSQL nos referimos únicamente a un tipo de bases de datos sino a diferentes soluciones dadas para almacenar datos los cuales las bases de datos relacionales no pueden gestionar su almacenamiento. Las bases de datos NoSQL son sistemas de almacenamiento de información que no cumplen con el esquema entidad-relación, no imponen una estructura de datos en forma de tablas y relaciones entre ellas, en ese sentido son más flexible.

El principio fundamental de esto gestores de datos es manejar el almacenamiento de datos que no poseen una estructura como tal, es por tal motivo que parten de la base en la que las "tablas" no existen como tal, sino que la información se almacena de forma distinta, generalmente como clave-valor, como una tabla en la que las columnas son dinámicas, pueden cambiar sin perder la agrupación de la información, es por esta razón que están pensadas para manipular enormes cantidades de información de manera muy rápida y variada. La información manejada allí es en base a formato de clave-valor (similar a tablas Hash), mapeo de columnas, documentos o Grafos.

Tipos de base de datos NoSQL

Clave-Valor: son las encargadas de almacenar cada elemento asignándolo a una llave única, lo que permite la recuperación de la información de manera muy rápida. De esta forma el tipo de contenido no es importante para la base de datos, solo la clave y el valor que tiene asociado. Son muy eficientes para lecturas y escrituras, además de que pueden escalar fácilmente particionado los valores de acuerdo a su clave; Dentro de estas bases de datos podemos encontrar a BigTable de Google, SimpleDB de Amazon, Cassandra, Hadoop, Riak, Voldemort y MemcacheDB entre otras.

Basada en Documentos: se encargan de almacenar la información en forma de documento (generalmente con una estructura simple como JSON o XML) , el cual el mismo está identificado con una clave única. El comportamiento es similar a las bases de datos clave-valor, pero con la diferencia que el valor es un fichero que puede ser entendido por el servidor. En este grupo se encuentra MongoDB y CouchDB entre las más importantes de este tipo.

Orientadas a Grafos: Son bases de datos que almacenan la información como grafos, donde las relaciones entre los nodos son lo más importante. Son muy útiles para representar información de redes sociales. Cabe destacar que las relaciones pueden tener atributos por lo se pueden ejecutar consultas directas a relaciones, en vez de a los nodos. Es importante señalar que este tipo de bases de datos sólo son útiles si la información a almacenar se puede representar fácilmente como una red. En este grupo tenemos a neo4j y otras de este estilo.

- Orientadas a Columnas: en este caso se guardan los valores en columnas en lugar de filas. Con esta característica se obtiene mucha velocidad en lecturas, ya que si se requiere consultar un número reducido de columnas, se ejecuta rápidamente sin embargo las tareas de escrituras no se ejecuta eficientemente. Es por esta razón que este tipo de soluciones es usado en aplicaciones con un índice bajo de escrituras pero que ejecuta muchas tareas de lecturas. En este grupo tenemos a Cassandra.

Arquitectura Orientada a Servicios

La Arquitectura Orientada a Servicios es un nuevo concepto de arquitectura de software que define la utilización de servicios para la creación de sistemas altamente escalables que reflejan el negocio de la organización, a su vez brinda una forma bien definida de exposición e invocación de servicios, lo cual facilita la interacción entre diferentes sistemas propios o de terceros.

Según Quispe (2011) la define como un sistema descrito en servicios (estos pueden ser componentes) y la composición entre estos (relaciones), las características que la conforma son las siguiente:

- Es una arquitectura basada en estándares.
- Los servicios son autónomos y granulares
- Los proveedores y consumidores se encuentran débilmente acoplados.

La Arquitectura Orientada a Servicio es una forma de arquitectura de sistemas distribuidos que se caracterizan por ofrecer un un marco de trabajo conceptual que permite a las organizaciones unir los objetivos de negocio con la infraestructura de TI integrando los datos y la lógica de negocio de sus sistemas separados.

Valera (2013) asegura que los componentes son la mejor forma de implementar servicios, pero el autor asegura que se debe entender que una aplicación correctamente basada en componentes, no necesariamente es una aplicación correctamente orientada a servicios.

Quispe (2011) asegura que un servicio en SOA, es una unidad de software con una funcionalidad mínima, que presenta las siguientes características:

- Interfaz definida o contrato de servicio: Descripción de como el servicio va ser usado desde cualquier servicio o programa: Nombre, parámetros, resultados y ubicación.

- **Reutilizable y/o componible con otros:** significa que pueda ser utilizado por más de una aplicación y/u otros servicios.

- **Desacoplado:** Que para prestar su funcionalidad dependa en lo mínimo de otro servicio.

Mecanismos de Recuperación utilizados para el manejo de Datos generados en Redes Sociales

Al tener un análisis de los conceptos teóricos que intervienen en la propuesta, a continuación se identifican los mecanismos de búsquedas y recuperación que se utilizan para tratar los contenidos generados en redes sociales. Mysore y otros (2013) confirman que los métodos más populares (ver figura 17), que se usan para recuperar datos generados en la web y social media son : crawler social, APIs Social , métodos de aceleración de contenido y web crawlers.

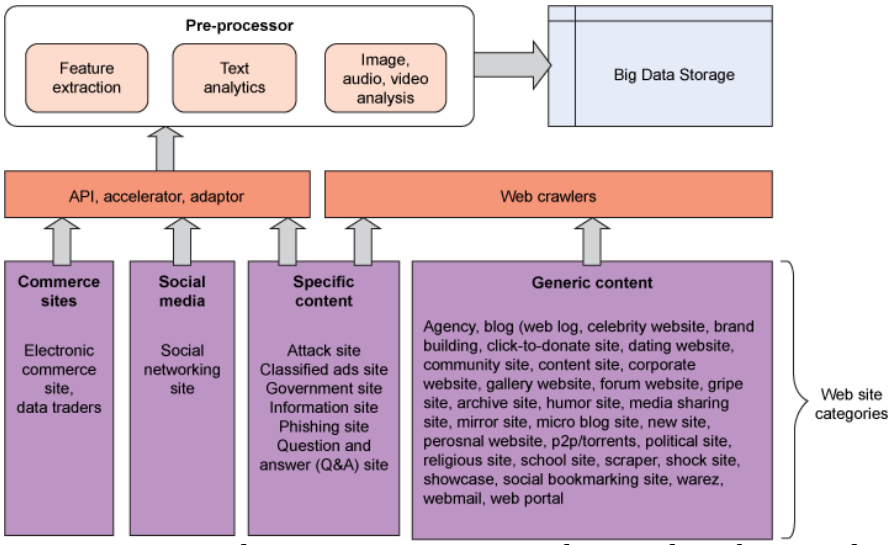


Figura 17. Mecanismos de acceso para recuperar datos en la web y social media.
Fuente: Mysore y otros (2013)

Como el ámbito del actual estudio es en detallar los métodos usados para tratar y capturar datos de redes sociales, a continuación se describe los comportamientos de los procedimiento de Crawler social y APIs Social:

Crawler social

Canali y otros (2011), afirma que el "crawling" (rastreo) es la forma más popular para la adquisición de datos en las redes sociales y consiste en consultar la información que esta publicada de manera disponible por los usuarios de la red social.

Este proceso lo ejecuta un componente denominado "crawler", el cual trabaja en conjunto con las API públicas que ofrecen algunas redes sociales para la lectura de la información.

El crawler social lo denomina los autores Canali y otros (2011) apoyándose en el estudio de Mislove (2007), como el rastreo que explora la estructura típica de una red social, que puede ser modelada como una dirigido grafo G (U, E), donde U es el conjunto de nodos (usuarios) y E es el conjunto de aristas (vínculos sociales entre usuarios). Cada nodo tiene enlaces salientes y enlaces entrantes.

Gjoka y otros (2010) asegura que el rastreo del grafo de una red social es un proceso iterativo que comienza a partir de un conjunto de usuarios iniciales y en el proceso se van descubriendo nuevos usuarios en cada paso. Los autores señalan que la configuración inicial del crawler debe ser compuesta por una lista de usuarios seleccionados al azar, porque a partir de múltiples ubicaciones aleatorias es una forma de mejorar el proceso de recolección de datos en términos de duración y de representatividad.

En la actualidad existen método matemáticos que ayudan a recorrer la estructura de un grafo según las condiciones mencionadas anteriormente, entre los cuales se encuentran Breadth-Search-First (BSF)(Búsqueda en anchura), Depth-First-

Search (DFS)(Búsqueda en profundidad), Forest Fire (FF) (Método de incendios forestales) and Snowball Sampling (SBS)(Muestreo de bola de nieve). Según los objetivos a seguir se selecciona el mecanismo apropiado que va a dirigir el orden en el cual serán visitados todos los nodos que conforman el grafo. Canali y otros (2011) asegura que el algoritmo BSF es el que se aplica para estos casos, debido a que se utiliza ampliamente en la literatura para la recolección de datos de redes sociales, esto basándose en los estudios de tipo muestral de Mislove (2007), Ahn (2007) y Wilson (2009).

A continuación, se presenta las siguientes propiedades definidas por Canali y otros (2011), los cuales comentan que deben estar presente en un crawler social, para ejecutar la exploración de datos en redes sociales:

- **Semillas :** Representa los nodos iníciales seleccionados al azar para empezar la exploración.

- **Selección del algoritmo:** Este va a establecer el orden de cómo serán visitados los nodos en el grafo.

- **El tamaño de los sub-grafos a explorar:** está relacionado con la exploración de la estructura típica de la red social, está relacionado a identificar propiedades tales como el grado del nodo, la distribución de los nodos entre otras.

API Social

Las APIs (Interfaces de programación de aplicaciones) son instrucciones y herramientas para la gestión de las interacciones entre diferentes software y pueden utilizarse para extraer automáticamente datos de medios de comunicación social. Taylor y Hobbs (2014) comenta que algunas plataformas de medios sociales proporcionan APIs gratis, solo si no imponen restricciones a la cantidad o el tipo de datos que se pueden acceder.

Las APIs de redes sociales igual que cualquier otra interfaz de programación

brindan al programador un conjuntos de funciones y procedimientos con el fin de ofrecer sus bibliotecas para ser utilizado por los sistemas externos que quieran interactuar con ellos.Tal como lo confirma Moral (2011) , las APIs son utilizada en el ámbito de la web 2.0 como pequeños sistemas para consultar bases de datos externas de forma automatizada, donde a través de ella se pueden construir nuevas webs usando información procedente de otras.

Sandoval y Hernández (2014) confirman que hoy en día existen muchas aplicaciones que están siendo diseñadas de tal forma que tengan la capacidad de establecer una conexión con una API social, esto con el fin de conocer la información que se generan en estos espacios sociales para generar procesos como: posicionamiento de marcas o servicios, procesos de minería de datos, inteligencia de negocio, entre otros .Los autores Sandoval y Hernández (2014), resaltan tres características comunes que presentan estos conjuntos de APIs en las distintas redes sociales:

- **Protocolo OAuth "Protocolo abierto de autorización" (Open Authentication):** Este protocolo permite que un usuario conceda acceso a un tercero (proveedor de servicio o aplicación) para que acceda a sus datos, sin tener que proporcionarle su usuario y contraseña. Esta conexión es la que permite el proceso de autenticación y autorización entre la API y los sistemas terceros interesado en autenticarse.

- **Generación de token por petición:** al momento en que el usuario concede permisos a la aplicación, la red social proporciona un "token" que deberá ser guardado por dicha aplicación para poder realizar peticiones en nombre del usuario, tales como leer información personal, intereses, contactos o publicar nueva información.

- **Interacción y solicitudes al servidor social:** La interacción entre las redes sociales y la aplicación es realizada mediante peticiones enviadas sobre el protocolo HTTPS. Dependiendo de la acción que se desee realizar pueden usarse peticiones de

tipo GET, POST, PUT o DELETE. Todas las peticiones, independientemente si son para leer, escribir, editar o eliminar información, deben de incluir el "token" de acceso por medio del cual se valida la petición y se acepta o se rechaza de acuerdo a los permisos otorgados por el usuario. Es recomendable que el "token" de acceso sea enviado dentro del encabezado de las peticiones, sin embargo, también puede ser enviado como parámetro en la URL, ya que es protegido por el mismo protocolo HTTPS.

Teniendo en cuenta las características principales que tiene una API social, a continuación se procede a mostrar en el cuadro 7, el resumen de todas las API social que actualmente son comúnmente utilizada, presentado por los autores Kanoulas y otros (2012) en su artículo denominado " D2.3 Standards, Software Interface Modules and APIs for inter-platform communication in Web 2.0 Social Media" (Normas D2.3, módulos de interfaz de software y APIs para la comunicación entre la plataforma en la Web 2.0 en Redes Sociales).

Cuadro 7
API de social media utilizadas frecuentemente

APIs	**Versión**	**Formato de Dato**	**Estilos Llamados**	**Protocolos Autorizados**	**Métodos de recuperación**
Blogger	Blogger API v3.0	JSON	REST, REST from Javascript, client libraries	OAuth 2.0 / API key	getByUrl listas
Wordpress	REST API	JSON	REST	OAuth 2.0 / API key	Get blog's posts Get likes for a post Get comments for a post
Facebook	Graph API	JSON	REST, REST from PHP, REST from	OAuth 2.0 / API key	Get comements

			JavaScript, REST from Android, REST from iOS SDK, FQL		Get Likes Get id Get locale Get languages Get bio Get birthday Get friends Get location Get relationship status
Twitter	REST API	JSON	REST, REST from ActionScript/Flash, REST from C++, REST from Clojure, REST from ColdFusion, REST from Erlang, REST from Java, REST from Javascript, REST from .NET, REST from Objective C / Cocoa, REST from Perl, REST from PHP, REST from Python, REST from Ruby, REST from Scala	OAuth 2.0 / API key	Get UserTimeLine Get retweets of a tweet

YouTube	2(version 3 exists as well, but is not used herein due to its experimental state)	XML, JSON	REST, REST from Java, REST from .NET, REST from PHP, REST from Python, REST from Objective –C, REST from Javascript	OAuth 2.0 / API key	Browsing with Categories and Keywords . Retrieving data for single video. Retrieve user profiles Retrieve comments.
Google+	3.0 (Previous 1.0, 2.0)	JSON	REST, REST from Javascript, REST from python, client libraries	OAuth 2.0 / API key	Search List get
LinkedIn	REST API	JSON, XML	REST from Java, REST from Javascript, REST fromPHP, REST from Python, REST from Objective C, REST from Ruby, REST from Clojure	OAuth 2.0 / API key	Get Status updates comments (STAT Updates) Get Status updates likes (STAT Updates) Get characteristics of a user's profile

Fuente: Kanoulas y otros (2012)

Elementos que intervienen el proceso de la construcción de un repositorio en el marco de redes sociales

Mysore y otros (2013), comentan que cuando se gestionan datos de tipo no estructurados, es esencial manejar tres procesos fundamentales: captura, transformación y almacenamiento. Es por lo anteriormente comentado que los autores destaca la importancia de definir bien el escenario en donde se están generando estos datos, es decir definir las fuentes de información a trabajar, ya que en algunas

estructuras de datos de este tipo, pueden requerir de procedimientos adicionales que son necesarios para lograr su consistencia y estandarización antes de procesar su almacenamiento. En la figura 18 se muestran los pasos a seguir detallan los procesos de captura y almacenamiento de los datos generados en la web y social media. Seguida de la figura anterior, se observa el cuadro 8 con distintos escenarios, en donde cada uno describe los distintos pasos que ejecutan dependiendo del tipo de dato (estructurado, no estructurado o semi-estructurado) a gestionar y el método de captura utilizado(crawler , APIs social).

Tomando la información anterior como base, cabe destacar entonces que para el caso de los datos generados en redes sociales, los elementos básicos que intervienen para lograr la construcción de un repositorio con esos datos, serian: un componente rastreador, un componente que ejecute la acción de estandarización y un repositorio que soporte almacenamiento de datos no estructurados.

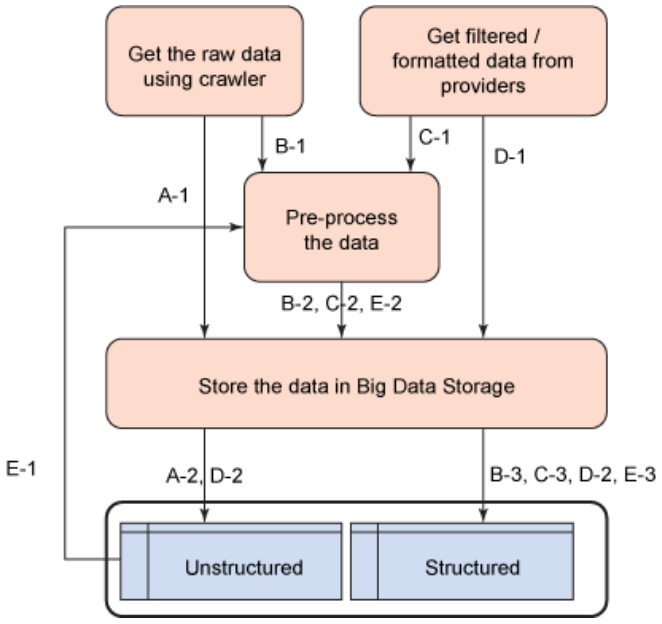


Figura 18. Pasos que detallan la captura y almacenamiento de los datos generados en la web y social media.
Fuente: Mysore y otros (2013**).**

Cuadro 8
Descripción de cada escenario para la gestión de datos generados en la web y social

Escenario A: Acceso a los medios web para los datos en el almacenamiento estructurado
Paso A- 1 . Un rastreador lee los datos en bruto.
Paso A- 2 . Los datos se almacenan en el almacenamiento no estructurado
Escenario B: Acceso a los medios Web de datos pre - proceso de almacenamiento estructurado
Paso B- 1 . El rastreador lee los datos en bruto.
Paso B- 2 . Estos datos se pre - procesado.
Paso B- 3 . Los datos se almacenan en el almacenamiento estructurado.
Escenario C: Acceso a los medios sociales para pre - proceso de datos no estructurados
Paso C-1. Datos de los proveedores pueden ser estructurados, en casos raros.
Paso C-2. Los datos son previamente procesados.
Paso C-3. Datos se almacenan en almacenamiento estructurado.
Escenario D: Acceso a los medios social para datos no estructurados o estructurados
Paso D- 1 . Los proveedores de datos proporcionan datos estructurados o no estructurados .
Paso D- 2 . Los datos se almacenan en el almacenamiento estructurado o no estructurado
.**Escenario E: Acceso a los medios web para pre - proceso de datos no estructurados**
Paso E-1. Los datos no estructurados, almacenados sin tratamiento previo, no pueden ser útiles si está en un formato estructurado.
Paso E-2. Los datos son previamente procesados.
Paso E-3. Se almacenan datos pre-procesados, estructurados en almacenamiento estructurado.

Fuente: Mysore y otros (2013).

Mapa Conceptual

A partir de la temática que aborda el dominio de la aplicación, en un mapa conceptual se puede representar la relación e interacción de los principales conceptos que intervienen en la construcción de un repositorio para almacenar datos generados en redes sociales. La figura 19 permite conceptualizar un modelo funcional de los procesos, lo cual ayuda a establecer relaciones de dependencia entre los conceptos y es útil para modelar el sistema.

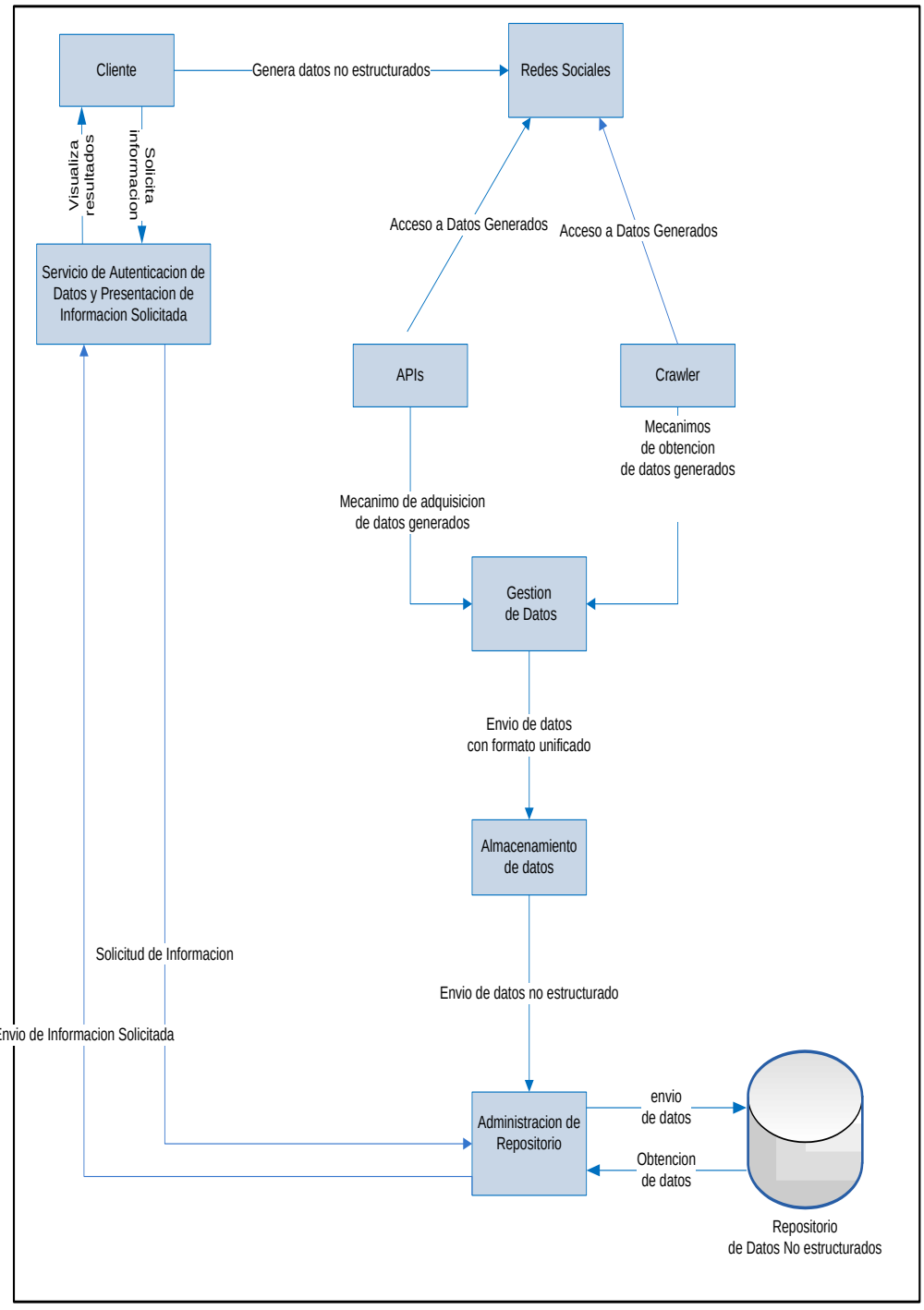


Figura 19. Mapa Conceptual de la arquitectura propuesta.
Fuente: El autor de la investigación

Especificación de Requisitos

El primer paso a seguir para el análisis de una propuesta de arquitectura de software es proveer de una especificación de requisitos; el cual se ha estructurado tomando como base algunas de las directrices dadas por el estándar "IEEE Recommended Practice for Software Requirements Specification ANSI/IEEE 830 1998". Debido a que no se trata de un documento individual dedicado únicamente a la especificación de requisitos, se ha incluido solamente aquellas partes consideradas útiles y necesarias para la correcta interpretación de los requisitos.

Propósito: El objeto de la especificación es definir de manera clara y precisa todas las funcionalidades y restricciones de la aplicación que se desea construir.

Conceptos y Términos: En este apartado se describirá brevemente algunos de los conceptos que puedan utilizarse a lo largo del documento y cuyo significado pueda resultar difuso o desconocido para el lector.

Términos utilizados como sinónimos:

El término de **aplicación, sistema, programa y solución propuesta** se usarán como sinónimos para hacer referencia al término de arquitectura propuesta con el objetivo de no escribir textos largos y evitar posibles redundancias en su uso.

El término **crawler social** se usará como sinónimo para referenciar al componente rastreador que interviene en el sistema.

El término **token**, se usará para representar las solicitudes acceso o de autorización que se envían a las interfaces de programación de aplicaciones suministradas por los servicios de redes sociales.

El término de **contenidos generados**, se usará para representar los contenidos digitales como video, fotos, comentarios, texto entre otros, que son creado publicados o compartido por el usuario en las diferentes redes sociales.

El término de **relaciones,** se usará como sinónimo para hacer referencia al término de lista de amigos o de contactos que contenga un perfil social específico.

El término de **operaciones,** se usará para representar las acciones de propagación que están disponibles en las diferentes redes sociales para calificar y compartir los contenidos digitales creados por los usuarios.

El término de **interfaz de consulta** se utilizará como sinónimo para representar la interfaz del usuario final y que también podrá llamarse **interfaz de búsqueda.**

Acrónimos:

HTML: Acrónimo inglés de Hypertext Markup Language, es decir, lenguaje de formato de documentos de hipertexto).

ERS Especificación de Requisitos Software.

API Social: Del inglés Application Programming Interface Social, es decir, son interfaces de programación de aplicaciones que facilitan la integración de los servicios de redes sociales con cualquier sistema o software a usar.

HTTP: Del inglés HyperText Transfer Protocol, es decir, protocolo de transferencia de hipertexto

oAUTH: Del inglés Open Authorization, es decir, Protocolo Abierto de Autorización.

NoSQL: Acrónimo inglés de Not Only SQL, es decir, base de datos no relacionales.

XML: Acrónimo inglés de eXtensible Markup Language, es decir, lenguaje de marcas extensible.

UGC: Acrónimo inglés de User Generated Content, es decir, Contenido Generado por el Usuario.

Diccionario de Actores

- **Usuario.** Cualquier persona que tener la posibilidad de registrarse y usar la interfaz de búsqueda, con la intención de consultar los datos que se encuentran almacenado en el repositorio de acuerdo a criterios o palabras claves introducidos por el mismo.
- **Administrador.** Es un tipo de usuario especializado de cargar los parámetros de configuración, como adecuar el sistema(repositorio de datos, plataforma de multiagente) y así como también activar los procesos de rastreo y extracción datos.
- **Agente Rastreador-Extractor** : Son los elementos automatizados (no humanos) que se encargaran de rastrear los perfiles de usuarios que estén registrados en el sistema y a su vez se encargaran de extraer los datos generados en esos perfiles, convirtiéndolos en formato único para su posterior almacenamiento.
- **El sistema** el cual invoca los procesos dentro del mismo.

Identificación de los Requisitos

Requisitos Funcionales

En este apartado se presentan los requisitos funcionales que deberán ser satisfechos por la arquitectura propuesta. Todos los requisitos aquí expuestos son esenciales, es decir, no sería funcional la arquitectura propuesta si algunos de estos requisitos faltase.

RF-1: La Arquitectura debe estar en la capacidad de acceder a varios sitios de redes sociales.

RF-2: La aplicación deberá ofrecer una interfaz donde el usuario pueda registrarse y autorizar a la aplicación de rastrear su perfil social, seleccionando así los sitios de redes sociales a monitorear.

RF-3: La arquitectura debe estar en la capacidad de rastrear los contenidos generados, relaciones (listas de contactos) y operaciones que se encuentren asociado a los perfiles de usuarios que estén previamente registrados en la aplicación.

RF-4: El sistema debe ser capaz de extraer la información previamente rastreada de los perfiles de usuarios que estén autorizados.

RF-5: La arquitectura debe estar en la capacidad de contar con un repositorio de datos en donde se almacenen todos los datos y perfiles de usuarios extraídos.

RF-6: La aplicación debe permitir el manejo de un formato estándar, para transformar los datos previamente extraídos para gestionar su posterior almacenamiento.

RF-7: Al momento de almacenar la información es necesario registrar los datos del perfil del usuario asociado como también su lista de contactos, de tal manera que se pueda llevar un histórico de información generada.

RF-8: La arquitectura deberá contar métodos de consultas, en donde le permita al usuario a través de una interfaz realizar búsquedas basados en palabras claves, y visualizar la información preliminarmente almacenada en el repositorio.

RF-9: Garantizar la seguridad de los datos y contenidos generados que compartan dichos usuarios al momento de registrarse, solamente van a ser almacenados en el repositorio autorizado.

RF-10: Garantizar el crecimiento exponencial de los contenidos digitales recolectados ahora y en el futuro.

Requisitos No Funcionales

RNF-1: La arquitectura propuesta estará basada en un sistema distribuido, sin restricciones tecnológicas notables.

RNF-2: Debe existir un componente para llegar a cada fuente de dato social, para ellos se utilizan APIs social o proveedores de datos social especifico, estos componentes varían por cada fuentes de datos.

RNF-3: El componente que va a ir a cada una de las fuentes de datos debe rastrear y extraer dos cosas: a) el perfil del usuario con su debida lista de amigos, b) contenido que genera el usuario.

RNF-4: Los componentes que se encarguen de rastrear y extraer los datos deben ejecutarse de manera paralela y sincrónica, con la finalidad de proveer simultáneamente el acceso y captura de los datos en las diferentes redes sociales.

RNF-5: Los componentes que se encarguen de manejar la gestión del almacenamiento de los datos deben ejecutarse de manera paralela, con la finalidad de proveer simultáneamente la consulta y el acceso a los datos que estén en el repositorio.

RNF-6: Se deberá definirse una estructura de datos para garantizar el almacenamiento uniforme de los datos previamente extraídos en el repositorio.

RNF-7: El diseño de la arquitectura deberá ofrecer modularidad para poder reemplazar partes del mismo con facilidad en un futuro.

RNF-8: Los datos almacenados en el repositorio han de ser siempre consistentes.

RNF-9: El sistema deberá ser tolerante a fallos de subsistemas externos.

RFN-10: El repositorio de datos no estructurados debe contar una infraestructura flexible, para preservar el contenido a largo plazo.

Modelo de Calidad

Muchos investigadores consideran que los requisitos no funcionales, expresados en términos de calidad, son cruciales para el diseño arquitectural, específicamente cuando las aplicaciones deben responder a situaciones críticas y a cambios en el ambiente, es por tal motivo se procederá a enmarcar los requisitos no funcionales, adoptando un modelo de calidad. Para tal fin, en esta investigación se empleará la Norma ISO-25010, por ser una de las más usadas y recomendadas, como se especifica en el cuadro 9.

Cuadro 9
Propiedades de Calidad Asociados a los Requisitos No Funcionales

Requisitos No Funcionales	**Propiedades de Calidad Asociada (Características de Calidad) ISO-25010**
• **RNF-1:** La arquitectura propuesta estará basada en un sistema distribuido, sin restricciones tecnológicas notables.	**Portabilidad:** Adaptable
• **RNF-2:** Debe existir un componente para llegar a cada fuente de dato social, para ellos se utilizan APIs social o proveedores de datos social especifico, estos componentes varían por cada fuentes	**Compatibilidad:** Interoperabilidad.

Requisitos No Funcionales	Propiedades de Calidad Asociada (Características de Calidad) ISO-25010
de datos.	
• **RNF-3*:*** La arquitectura debe adaptarse para acceder a diferentes tipos de documentos, que se originan en las redes sociales.	**Fiabilidad:** Madurez
• **RNF-4:** Los componentes que se encarguen de rastrear y extraer los datos en las diferentes redes sociales, deben ejecutarse de manera paralela y sincrónica, con la finalidad de proveer simultáneamente la captura de la información en las diferente fuentes sociales.	**Fiabilidad:** Disponibilidad
• **RNF-5:** Los componentes que se encarguen de manejar la gestión del almacenamiento de los datos deben ejecutarse de manera paralela, con la finalidad de proveer simultáneamente la consulta y el acceso a los datos que estén en el repositorio.	**Funcionalidad Apropiada:** Completa. **Eficiencia:** Tiempo de Repuesta.
RNF-6: Se deberá definirse una structura de datos para garantizar el lmacenamiento uniforme de los datos reviamente extraídos en el repositorio.	**Compatibilidad:** Coexistencia.
• **RNF-7:** El diseño de la arquitectura	**Mantenibilidad:** Modularidad

Requisitos No Funcionales	Propiedades de Calidad Asociada (Características de Calidad) ISO-25010
deberá ofrecer modularidad para poder reemplazar partes del mismo con facilidad en un futuro.	
• **RNF-8:** Los datos almacenados en el repositorio han de ser siempre consistentes.	**Seguridad:** Integridad
• **RNF-9:** El sistema deberá ser tolerante a fallos de subsistemas externos. • **RFN-10:** El repositorio de datos no estructurados debe contar una infraestructura flexible, para preservar el contenido a largo plazo.	**Fiabilidad:** Tolerancia a Fallos. **Mantenibilidad:** Capacidad de ser modificado. **Seguridad:** Confidencialidad **Portabilidad:** Reemplazable.

Fuente: El autor de la investigación

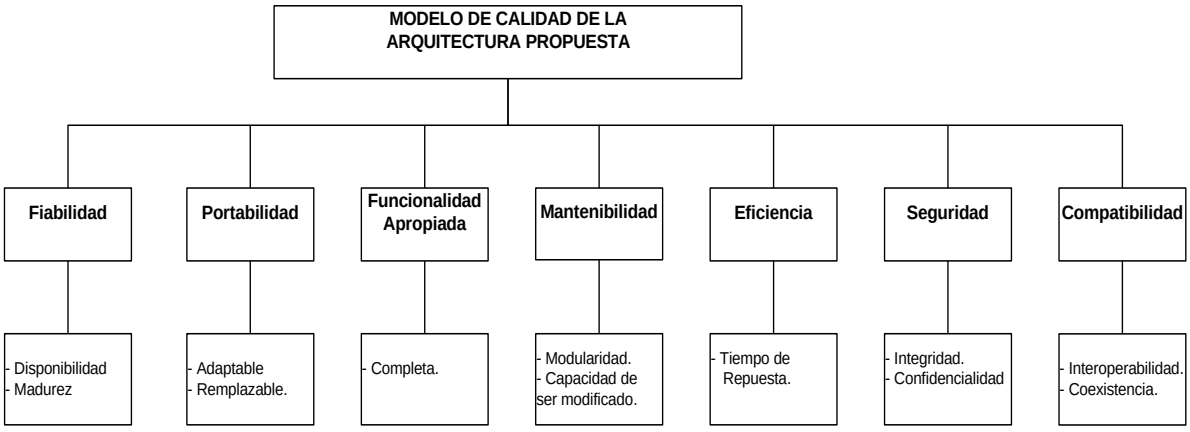


Figura 20 Modelo de Calidad de Arquitectura Propuesta.
Fuente El autor de la investigación

Atributos de Calidad de la solución propuesta

Se establecen atributos y métricas de calidad que son tenidos en cuenta para el inicio del proceso de evaluación del modelo de calidad de Arquitectura Propuesta.

Cuadro 10
Atributos de Calidad Asociados a los Requisitos No Funcionales

Requisitos No Funcionales	**Propiedades de Calidad Asociada (Características de Calidad)ISO-25010**	**Atributos de Calidad (Características)**
• **RNF-1:** La arquitectura propuesta estará basada en un sistema distribuido, sin restricciones tecnológicas notables.	**Portabilidad:** Adaptable	**Atributo:** El sistema estará disponible para usarse en cualquier plataforma tecnología **Métrica** booleano.
• **RNF-2:** Debe existir un componente para llegar a cada fuente de dato social, para ellos se utilizan APIs social o proveedores de datos social especifico, estos componentes varían por cada fuentes de datos.	**Compatibilidad:** Interoperabilidad.	**Atributo**: Presencia de componentes adaptables por cada fuente social. **Métrica:** Un número en el rango [1..10].
• **RNF-3*:*** La arquitectura debe adaptarse para acceder a diferentes tipos de documentos, que se originan en las redes sociales.	**Fiabilidad:** Madurez	**Atributo:** Presencia de mecanismos que permitan realizar las operaciones de manera rápida y sin fallos provisto. **Métrica:**Booleano.
• **RNF-4:** Los componentes que	**Fiabilidad:**	**Atributo:** Presencia de

Requisitos No Funcionales	Propiedades de Calidad Asociada (Características de Calidad)ISO-25010	Atributos de Calidad (Características)
se encarguen de rastrear y extraer los datos en las diferentes redes sociales, deben ejecutarse de manera paralela y sincrónica, con la finalidad de proveer simultáneamente la captura de la información en las diferentes fuentes sociales.	Disponibilidad	mecanismos que permitan ejecutar operaciones distribuidas y sincrónicas, con el fin de abarcar de manera simultánea el mayor número de redes sociales **Métrica:** Booleano.
• **RNF-5:** Los componentes que se encarguen de manejar la gestión del almacenamiento de los datos deben ejecutarse de manera paralela, con la finalidad de proveer simultáneamente la consulta y el acceso a los datos que estén en el repositorio.	**Funcionalidad** **Apropiada:** Completa. **Eficiencia:** Tiempo de Repuesta.	**Atributo:** Tiempo que tardan las operaciones en ejecutarse. **Métrica:** Un número en el rango [1..10].
• **RNF-6:** Se deberá definirse una estructura de datos para garantizar el almacenamiento uniforme de los datos previamente extraídos en el repositorio.	**Compatibilidad:** Coexistencia.	**Atributo:** Presencia de formato o estructura de datos estándar que sea compatible para gestionar almacenamiento. **Métrica:** Boolean
• **RNF-7:** El diseño de la arquitectura deberá ofrecer	**Mantenibilidad:** Modularidad	**Atributo:** Presencia de mecanismos que permita

Requisitos No Funcionales	Propiedades de Calidad Asociada (Características de Calidad)ISO-25010	Atributos de Calidad (Características)
modularidad para poder reemplazar partes del mismo con facilidad en un futuro.		realizar cambios en los algoritmos y métodos que compone la solución, sin perjudicar la funcionalidad que esta ya posea. **Métrica:** un número en el rango [1..10].
• **RNF-8:** Los datos almacenados en el repositorio han de ser siempre consistentes.	**Seguridad:** Integridad	**Atributo:** Presencia de mecanismos que validen que los datos obtenidos están completos y en el formato estándar deseado. **Métrica:** Booleano.
• **RNF-9:** El sistema deberá ser tolerante a fallos de subsistemas externos.	**Fiabilidad:** Tolerancia a Fallos. **Mantenibilidad:** Capacidad de ser modificado.	**Atributo:** Presencia de un Mecanismo de continuidad por parte del el sistema en caso de que los sistemas externos fallen. **Métrica:**Booleano.
• **RFN-10:** El repositorio de	**Seguridad:**	**Atributo:**

Requisitos No Funcionales	Propiedades de Calidad Asociada (Características de Calidad)ISO-25010	Atributos de Calidad (Características)
datos no estructurados debe contar una infraestructura flexible, para preservar el contenido a largo plazo.	Confidencialidad **Portabilidad:** Reemplazable.	Capacidad del repositorio para almacenar los contenidos gestionados que se generan en la mayoría de las redes sociales. **Métrica:** un número en el rango [0..10].

Fuente: El autor de la investigación

Diagrama de Casos de Usos

A partir de la información recopilada, elicitación de los requisitos, y las funcionalidades previamente descritas; surgieron los siguientes procesos que se tomarán en cuenta para el diseño de la arquitectura, por lo tanto el diagrama de casos de uso final queda como en la figura siguiente:

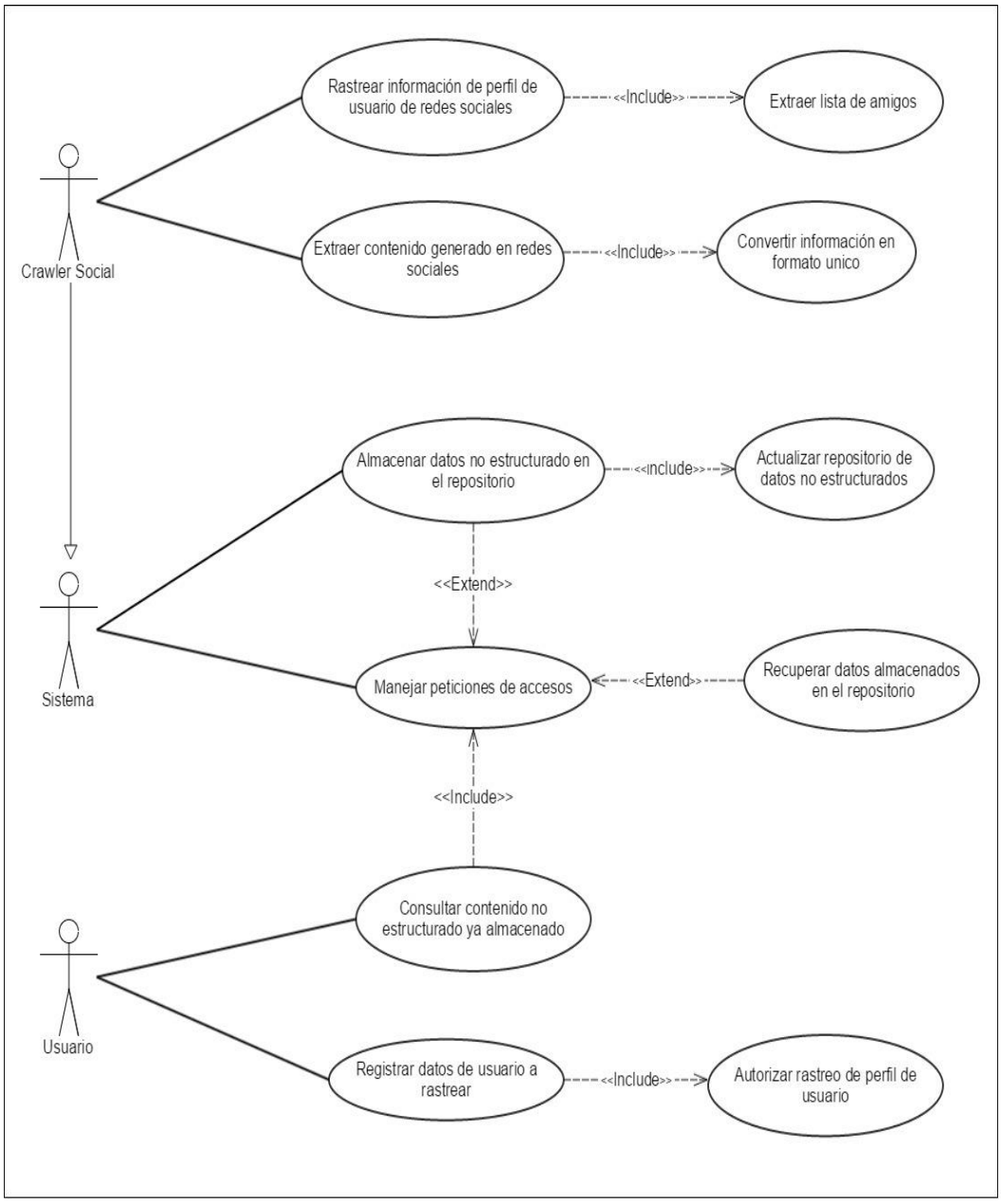


Figura 21 Diagrama de Casos de Usos
Fuente: Autor de la Investigación

Cuadro 11
Caso de Uso: Registrar Datos de Usuario a Rastrear

Nombre del Caso de Uso: Registrar Datos de Usuario a Rastrear		**Ref. 1**
Descripción: Permite al usuario de registrar sus datos de sección, como también autorizar al sistema el rastreo de sus diferentes perfiles sociales que tenga registrado.		
Pre-condición: Ninguna		
	Acción del Actor	***Acción del Sistema***
Curso Normal	1. El usuario entra a la interfaz del sistema y rellena los datos de sesión como: usuario y clave. 3.Al recibir la confirmación por parte del sistema, debe autorizar el rastreo de su información, indicando las redes sociales en donde tenga perfiles registrado	2. El sistema valida los datos y envía la confirmación al usuario. 4. El sistema recibe la autorización por parte del usuario y guarda las cuentas sociales asociadas al usuario.
Curso Alterno		2. El sistema valida los datos, si falta información, muestra un error a la interfaz, informando sobre la falta de datos.
Post-condición: Usuario registrado en base de datos.		

Cuadro 12
Caso de Uso: Autorizar Rastreo de Perfil de Usuario

Nombre del Caso de Uso: Autorizar Rastreo de Perfil de Usuario	**Ref. 2**
Descripción: El usuario autoriza al sistema, el rastreo de los perfiles sociales que tenga asociado	
Pre-condición: El usuario debe estar registrado previamente en el sistema	

Nombre del Caso de Uso: Autorizar Rastreo de Perfil de Usuario		**Ref. 2**
	Acción del Actor	***Acción del Sistema***
Curso Normal	1. El usuario al recibir la confirmación por parte del sistema, debe autorizar el rastreo de su información, indicando las redes sociales en donde tenga perfiles registrado	2. El sistema recibe la autorización por parte del usuario y guarda las cuentas sociales asociadas al usuario.
Curso Alterno		
Post-condición: perfiles de redes sociales asociadas al usuario autorizadas a ser rastreada.		

Cuadro 13

Caso de Uso: Consultar Contenido No Estructurado

Nombre del Caso de Uso: Consultar Contenido No Estructurado		**Ref. 3**
Descripción: Permite al usuario consultar información que fue publicada en las distintas redes sociales.		
Pre-condición: El usuario tiene que estar previamente registrado en el sistema y haber autorizado al sistema de rastrear sus perfiles asociados.		
	Acción del Actor	***Acción del Sistema***
Curso Normal	1. El usuario entra a la interfaz del sistema y indica que información quiere consultar.	2. El sistema valida los datos de sesión del usuario. 3. Si los datos de sesión son correcto, envía la petición al servicio que se encarga de consultar la información que se encuentra almacenada en el repositorio de datos no estructurados.

Nombre del Caso de Uso: Consultar Contenido No Estructurado		**Ref. 3**
		4. El servicio de consulta de información se conecta inmediatamente con el servicio de manejo de peticiones de acceso del repositorio. 5. Si la información solicitada se encuentra en el repositorio, el servicio de consulta de información obtiene el contenido solicitado y lo envía a la interfaz de usuario.
	6. El usuario recibe la información y visualiza el resultado de la búsqueda.	
Curso Alterno		2. El sistema valida los datos de sesión, si falla, muestra un error a la interfaz, informando sobre el error de sesión. 5. Si la información solicitada no se encuentra en el repositorio se muestra un error indicando que no se encontraron registros asociados a la búsqueda.
Post-condición: Obtención de información no estructurada proveniente de redes sociales previamente almacenada		

Cuadro 14
Caso de Uso: Manejar Peticiones de Acceso

Nombre del Caso de Uso: Manejar Peticiones de Acceso	**Ref. 4**
Descripción: Proceso que se encarga de gestionar las peticiones de acceso al repositorio de datos no estructurados.	
Pre-condición: Peticiones de acceso generadas por los procesos de consulta o almacenamiento de datos.	
	Acción del Sistema
Curso Normal	1. Recibe una petición de acceso por parte de un proceso de consulta de información o de un proceso de almacenamiento de información 2. Seguidamente se conecta con el conector NoSQL correspondiente para indexar al repositorio de datos no estructurado usando el modelo map reduce. 3. Si la petición gestionada proviene por un proceso de consulta, verifica si la información solicitada se encuentra almacenada, en caso de ser afirmativo envía un mensaje de repuesta en conjunto con la información en formato especifico para su posterior visualización. Si la petición proviene de un proceso de almacenamiento, recibe la información y la almacena en el repositorio.
Curso Alterno	
Post-Condición: Tareas de indexación al repositorio no estructurado.	

Cuadro 15
Caso de Uso: Almacenar Datos en el repositorio de Datos

Nombre del Caso de Uso: Almacenar Datos en el repositorio de Datos	**Ref. 5**
Descripción: El sistema toma los datos obtenidos del rastreo y los envía al repositorio de datos no estructurados.	
Pre-condición: Debe al menos estar activo un agente rastreador de contenido de redes	

Nombre del Caso de Uso: Almacenar Datos en el repositorio de Datos	**Ref. 5**
sociales	
	Acción del Sistema
Curso Normal	1. Recibe la petición de parte de un agente rastreador. 2. Obtiene la información adquirida por el agente rastreador como son el perfil de usuario y el contenido generado. 3. El sistema transforma el contenido obtenido en un formato JSON estándar para su posterior almacenamiento. 4. Envía una solicitud de tipo entrada, al servicio de manejo de peticiones de acceso del repositorio. 5. El servicio de manejo de peticiones de acceso, autoriza al sistema la entrada al repositorio. 6. El sistema accede al repositorio de datos no estructurados y verifica si la información obtenida, ya se encuentra guardada anteriormente. 7. En caso de que la información obtenida no se encuentre almacenada, el sistema procede a guardarla en el repositorio de datos no estructurados de la siguiente forma: primero verifica si el perfil del usuario existe en la base de datos de grafo, en caso contrario crea el nodo asignando las propiedades de ID único y nombre, este ultimo tomara el valor del nombre del usuario, seguidamente el sistema procesara el almacenamiento del contenido extraído, que corresponda cada nodo. 8.El sistema guardara cada contenido extraído, en una base de datos de tipo documento, por lo cual cada documento contendrá los siguientes datos: id del documento, nodo de usuario, fecha de extracción y tipo de contenido.
Curso Alterno	7. Si la información se encuentra ya almacenada, se procede a ejecutar el proceso de actualizar repositorio de datos no estructurados.
Post-Condición: Nuevos documentos ingresados al repositorio y actualizados en el caso de que se encuentre previamente guardado.	

Cuadro 16
Caso de Uso: Actualizar Repositorio de Datos No Estructurado

Nombre del Caso de Uso: Actualizar Repositorio de Datos No Estructurado	**Ref. 6**
Descripción: Proceso que comprende el conjunto de tareas que intervienen para la actualización del repositorio de datos no estructurados.	
Pre-condición: Debe al menos estar activo un agente rastreador de contenido de redes sociales, y el repositorio tener información previamente almacenada.	
	Acción del Sistema
Curso Normal	1. Recibe la petición de parte de un agente rastreador. 2. Obtiene la información adquirida por el agente rastreador como son el perfil de usuario y el contenido generado. 3. Envía una solicitud de tipo entrada, al servicio de manejo de peticiones de acceso del repositorio. 4. El servicio de manejo de peticiones de acceso, autoriza al sistema la entrada al repositorio. 5. El sistema accede al repositorio de datos no estructurados y verifica si la información obtenida, ya se encuentra guardada anteriormente. 6. En caso de que la información se encuentre almacenada, a dirigirse a la base de datos de documento y consulta según el id del documento. 7. El sistema al conseguir dicho documento, procede a sobrescribirlo con la información obtenida. Aquí el sistema puede modificar las propiedades del documento.
Curso Alterno	6. En caso de que la información no se encuentre almacenada, el sistema debe ejecutar el proceso de almacenar datos no estructurados en el repositorio.
Post-Condición: Documentos actualizados que se encuentre previamente guardado.	

Cuadro 17
Caso de Uso: Actualizar Repositorio de Datos No Estructurados

Nombre del Caso de Uso: *Actualizar Repositorio de Datos No Estructurados*	**Ref. 7**
Descripción: El sistema toma los datos obtenidos del rastreo y los envía al repositorio, si estos ya se encuentran previamente ya guardado realiza una actualización sobre lo mismos.	
Pre-condición: Debe al menos estar activo un agente rastreador de contenido de redes sociales	
	Acción del Sistema
Curso Normal	1. Recibe la petición de parte de un agente rastreador. 2. Obtiene la información adquirida por el agente rastreador y la transforma en un formato estándar para su posterior almacenamiento. 3. Envía una solicitud de entrada al servicio de manejo de peticiones de acceso del repositorio. 4. El servicio de manejo de peticiones de acceso autoriza la entrada al repositorio . 6. El sistema accede al repositorio de datos no estructurados y verifica si la información obtenida se encuentra ya guardada anteriormente, en caso de ser cierto sobrescribe los datos que se encuentran almacenados.
Post-Condición: Nuevos documentos actualizados en el repositorio.	

Cuadro 18
Caso de Uso: Rastrear Información de Perfil de usuario de redes sociales

Nombre del Caso de Uso: *Rastrear Información de Perfil de usuario de redes sociales*	**Ref. 8**
Descripción: La plataforma de agentes, activa en tiempo real el rastreo de los contenidos generados por los distintos perfiles sociales que pertenezcan a un usuario que se encuentre anteriormente registrado en el sistema.	
Pre-condición: Debe estar al menos un usuario registrado y adicional haber autorizado al sistema el rastreo de los distintos perfiles sociales que tenga asociado. Los agentes rastreadores estar inicializados.	

Nombre del Caso de Uso: *Rastrear Información de Perfil de usuario de redes sociales*	**Ref. 8**
	Acción del Sistema
Curso Normal	1. Selecciona al azar el perfil social a rastrear, el cual lo selecciona del grupo de usuarios que se encuentren registrados en el sistema y estén marcados como autorizados para rastrear. 2. Accede al perfil social a través del API de la red social correspondiente. 3. Del perfil de usuario obtiene el numero de contactos que se encuentren asociado a ese perfil. 4. Obtiene el grado del grafo, en donde determina el numero de nodos a rastrear y el conjunto de aristas, que representa los enlaces entrantes y salientes de cada nodo. 5. Activa el algoritmo Breadth-Search-First (BSF), el cual permitirá dirigir el orden en el cual serán visitados todos los nodos que conforman el grafo. 6. Ejecuta el proceso iterativo de rastreo en cada nodo, aplicando el algoritmo (BSF). 7. Al visitar cada nodo, envía una solicitud al agente extractor para iniciar el proceso de extracción del contenido generado.
Curso Alterno	
Post-Condición: Información rastreada generada en una red social especifica.	

Cuadro 19
Caso de Uso: Extraer contenido generados en redes sociales

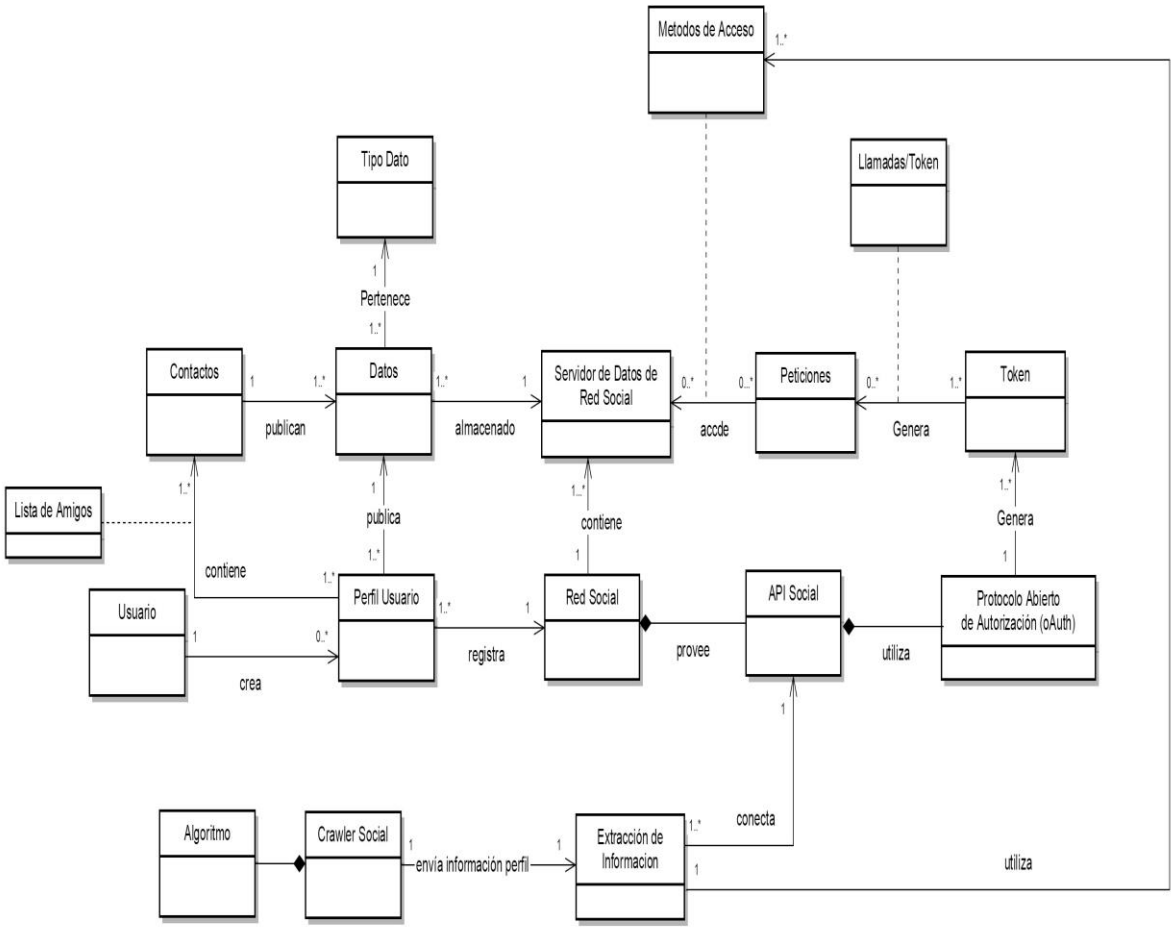
Nombre del Caso de Uso: Extraer contenido generados en redes sociales	**Ref. 9**
Descripción: Proceso que comprende el conjunto de tareas que intervienen para la extracción de información que se generan en los perfiles sociales que fueron rastreados previamente.	
Pre-condición: Debe al menos estar inicializado un agente rastreador de contenido de redes sociales, y tener perfiles sociales previamente rastreados.	

Nombre del Caso de Uso: Extraer contenido generados en redes sociales		**Ref. 9**
	Acción del Sistema	
Curso Normal	1. Recibe una confirmación de extracción por parte de un agente rastreador. 2. Se conecta a la API de red social correspondiente para verificar si tiene autorización para accesar al nodo a visitar. 3. Al recibir la confirmación por parte de la API, se ubica en el nodo rastreado y verifica si tiene contenido generado a sustraer. 4. Si tiene presenta contenido a extraer, se detalla el tipo y se inicia la extracción del mismo. 5. Extrae el contenido de acuerdo al tipo (documento, audio, imagen). 6. Convierte la información obtenida en un documento estándar en formato JSON, para su posterior almacenamiento. 7. Por cada documento estándar, se indica los siguientes datos como: id del documento, nodo perteneciente, fecha de extracción. 8. Realiza una solicitud para los servicios correspondiente de almacenamiento para gestionar la inclusión de la información extraída en el repositorio de datos no estructurados. 9. Al terminar de visitar todos los nodos, obtiene la lista de amigos del perfil social y datos principales del perfil, para incluirlo en la bd de grafo.	
Curso Alterno	.	
Post-Condición: Contenidos extraídos generados en perfiles de usuario en redes sociales.		

Modelo del Dominio

Como parte importante de modelar el dominio del problema; una vez que se conocen los conceptos involucrados y sus relaciones se puede hacer uso de artefactos tales como el diagrama de clase, que ayudó establecer las asociaciones de los recursos

u objetos a guardar; en este caso el recurso a almacenar es los contenidos generados por los usuarios en las redes sociales; en este sentido, se revisaron los conceptos y propiedades que caracterizan a este tipo de contenido; esto ayudó a diseñar la estructura que debe presentar el documento a guardar en el almacén de datos. El diagrama se describe en la figura 22.



Fase II. Determinar las Característica de un Repositorio de datos no estructurado, para almacenar datos proveniente de redes sociales.

La solución propuesta manejara dos base de datos en paralelo basado en tecnología NoSQL, la primera base de datos manejara un almacenamiento en forma de grafo y la segunda base de datos se encargara de almacenar documentos, el primero va a contener el grafo rastreado y extraído asociado a un perfil de usuario en especifico con la finalidad de reflejar las relaciones o lista de amigos que tenga este.

El segundo será una base de datos basada en documentos con la finalidad de guardar en un formato único todos los contenidos generados por cada nodo que fue rastreado. Katsov (2012) confirma que las bases de datos orientada a grafo están relacionadas con bases de datos documentales porque muchas implementaciones permiten un modelo de un valor como un plano o documento.

A continuación se presenta las característica que presenta cada base de datos con la finalidad de gestionar el almacenamiento de los contenidos generados por los usuarios en las redes sociales. Seguidamente se indicara que técnica de modelado de NoSQL se aplica para especificar el modelo de indexación para consultar la información almacenada.

Base de Datos orientada a Grafo

Katsov (2012) asegura que las bases de datos orientadas a grafos como Neo4j son excepcionalmente buena, específicamente para explorar los alrededores de un determinado nodo o las relaciones entre dos o varios nodos. Katsov (2012) asegura que el procesamiento de grandes grafo no es muy eficiente, porque las bases de datos de grafo de propósito general generalmente escalan bien. Es por esto que recomienda para el procesamiento para grafo distribuido aplicar MapReduce.

Base de datos orientada a Documentos

Tomando como base la comparación presentada por Fowler y Sadalage (2013) entre Oracle y MongoDB, a continuación el cuadro 20, se presenta una comparación entre una base de datos relacional y una base de datos NoSQL orientada a documento con el fin de presentar las propiedades que tendrá el repositorio orientado a documentos en la arquitectura propuesta:

Cuadro 20
Comparación de atributos entre base de datos relacional y base de datos tipo documentos

Base de Datos Relacional	Base de Datos NoSQL orientado a Documentos
database instance	database instance
schema	database
table	collection
row	document
rowid	_id

Fuente: Fowler y Sadalage (2013)

Para especificar la estructura del documento a guardar, se tomaron como base las técnicas de modelados de NoSQL mencionadas por Katsov (2012) y reforzadas por Fowler y Sadalage (2013), estas técnicas son las de agregados y claves enumeradas. Tomando como base el modelo de dominio planteado, en el cuadro 21 se muestra los atributos que conformara la estructura del documento a almacenar:

Cuadro 21

Atributos que conforman la estructura del documento a almacenar

Nombre	Descripción	Tipo
id	Identificador del documento	Integer
name	Nombre del documento	string
ExtracionTime	Fecha de extracción	Date
NodoUser	Usuario que genera la información, se guarda su id, nombre.	Array
Content	Contenido Extraído	byte
ContenType	Tipo de contenido, si el contenido es un texto, audio, imagen,	Array
protocol	Tipo de Protocolo	String
titleContent	Titulo del contenido	String
text	Descripción del Contenido.	String
headers	Guarda la Información Encabezada de la conexión	Array
metadata	Guarda la metadata del contenido.	Array

Fuente: Autor de la investigación

Los autores Fowler y Sadalage (2013), señalan que cuando se aplican técnicas de agregados a los datos, se tiene que especificar cómo se va a leer los datos. En la propuesta, se va trabajar con una base de datos orientada a documentos, en donde cada documento tiene asociado una clave única, la estructura de el objeto documento estará conformado de la siguiente manera: los datos del usuario, la red social y las características propias del contenido almacenado, va estar representado por la propiedad "Value", la identificación del documento va estar reflejada a través de la propiedad "Key" (clave), por lo tanto cuando el usuario quiera consultar un documento en especifico, tendrían que especificar el identificador único del objeto

para el acceso al contenido asociado, la figura 23 muestra con detalle lo anteriormente explicado:

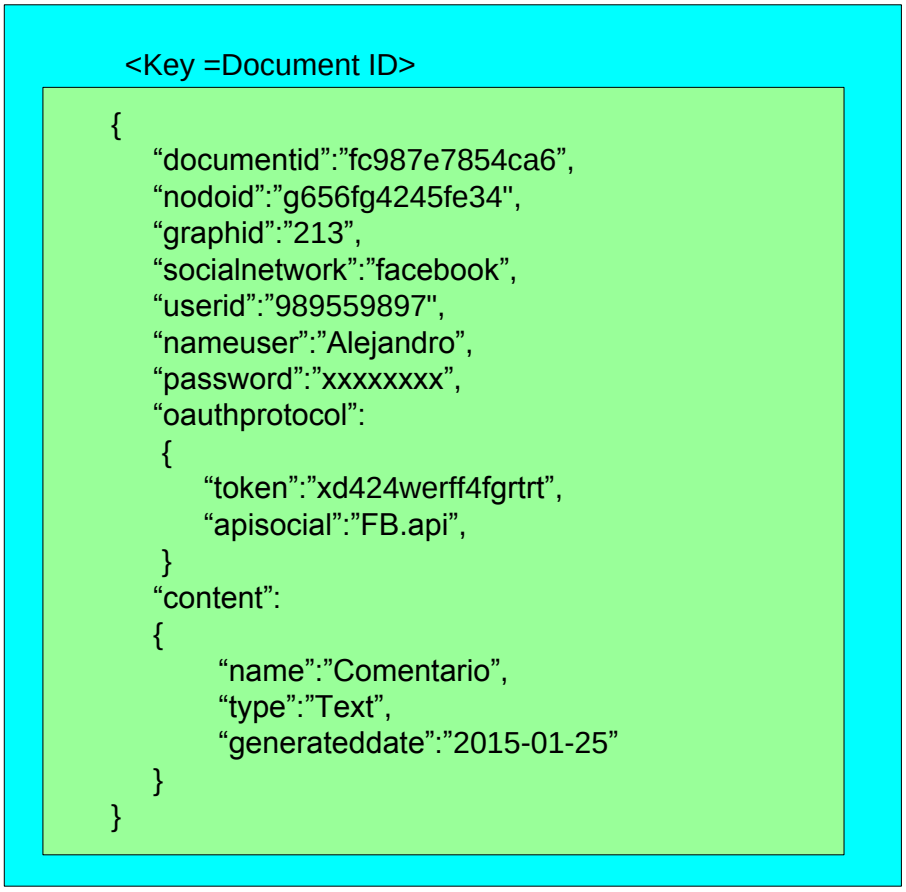


Figura 23. Almacenamiento de un documento.
Fuente: El autor de la investigación

Con este procedimiento, los autores Fowler y Sadalage (2013) asegura que permite la optimización de lectura, ya que la desnormalización de los datos permite un rápido acceso a los datos que interesan al usuario, permitiendo así entonces que procesos externos como análisis o técnicas de ETL de minería de datos, reduzca sus tiempo consultas sobre los datos almacenados.

Fase III Diseño de la Arquitectura Propuesta

Estructura de la Arquitectura Propuesta

Heller y Sun (2012) asegura que las capacidades fundamentales que debe tener una arquitectura Big Data son las siguientes: almacenamiento, administración, procesamiento, integración de datos y análisis estadísticos. Para el alcance de la arquitectura propuesta, se tomó como base el modelo propuesto por Mysore y otros (2013), considerando las funcionalidades y características de los procesos descritos por Boukhanovsky y Semenov (2011), esto permitió definir los cuatros procesos principales que cubre la solución: capturar, extraer, almacenar y consultar la información ya almacenada. Para mostrar de manera ordenada, estandarizada y independiente los componentes de software que intervienen en la solución, la arquitectura propuesta será descrita por capas, estas son : capa adquisición de datos y capa de almacenamiento.

La "capa adquisición de datos", está conformada por las APIs de cada red social y algunos componentes que actúan como crawler para aquellas fuentes sociales que no cuenta con proveedor de datos, ambos mecanismos representan la forma de obtención y recuperación de los contenidos generados por los usuarios en los distintos medios sociales, dicha capa está representada en un sistema multiagente, en el cual cada agente será independiente del uno del otro, es decir cada agente cuenta con una interface de transformación, el cual se encarga de exportar el dato previamente obtenido de la redes sociales en un formato estandarizado. La segunda capa llamada "Capa de Almacenamiento" se encarga de gestionar y almacenar los datos previamente extraídos, la misma está basada en un sistema orientado a servicios el cual solamente se encargara de las funciones almacenar y consultar. Dicha capa es

orientado a servicio debido a que cuenta con un grupo de servicios los cuales se encarga de manejar la persistencia del repositorio.

A continuación se describirá con detalle los elementos que intervienen en cada capa.

Capa Adquisición de datos

La capa de recolección y extracción de datos estará representada por un sistema multi-agente el cual estará conformado por un grupo de agentes inteligentes los cuales se comunicaran entre sí para ejecutar la recolección y el tratamiento de datos de forma paralela y masiva en las distintas redes sociales.

En esta capa se contara con los siguientes agentes para llevar a cabo las acciones anteriormente mencionadas:

Agente Activador

Es activado por el usuario administrador de sistema y lo cual desencadena la acción de rastrear los datos del sistema a través de la creación dinámica de toda los demás agentes que conforman la plataforma.

Agente Rastreador

Es el encargo de rastrear los perfiles de usuarios autorizados, está conformado por dos componente: el primero contiene el crawler a usar en donde implícitamente contiene el algoritmo de rastreo a activar y el segundo representa una memoria cache en donde se almacena de manera temporal el grafo de las red con sus respectivo nodos. Al tener dicho grafo se comunica a través través de la estructura de mensaje FIPA con el agente Extractor-Convertidor para que inicie la acción de extracción.

Agente Extractor-Convertidor

Es el encargado de recibir los eventos y la información de los agentes rastreadores cliente en cada fuente de dato, al momento de recibir el grafo asociado al

perfil usuario rastreado, genera un mensaje de repuesta de token de acceso con la finalidad de extraer los datos y contenido generados del perfil esto utilizando las APIs social de cada red, al momento de extraer los datos y relaciones, pasa esta información al componente convertidor en donde los transforma en un formato único para su posterior almacenamiento.

Agente Administrador

Es el agente responsable de gestionar la operatividad de la Plataforma de multiagentes, como creación de agentes, la eliminación de agentes, registro de agentes.

Este agente tendrá el rol de la coordinación de proceso de invocación de los agentes rastreador y extractor-convertidor a través de la creación threads (hilos) para reforzar el trabajo multiparalelo y obtener el máximo rendimiento de los recursos del servidor.

Agente mediador

Es el agente responsable de enviar los documentos a la capa de almacenamiento ya convertidos por el Agente extractor- convertidor, cuenta con un componente denominado facilitador el cual dirige las peticiones de los documentos enviar a los servicios que se encuentran en la capa de almacenamiento

Agente director

Es el agente responsable de mantener una lista precisa, completa y oportuna de todos los agentes del sistema, todos los agentes deben mantenerlo al tanto de las actualizaciones para el correcto registro de los agentes.

La comunicación entre los agentes se realiza a través del estándar FIPA-ACL-SL (lenguaje de comunicación de agentes semántico). Tal como lo comenta

Moros(2013) el contenido de los mensajes se puede procesar en cualquier lenguaje de dominio, el cual FIPA-ACL-SL ha sido enriquecido semánticamente en la forma de pre y post condiciones en cada uno de estos mensajes. Moros (2013) afirma que dentro del enfoque de comunicación entre agentes, la cooperación es realizada vía ACL (lenguaje de comunicación de agentes), el contenido del lenguaje y la ontología

la cual identifica el grupo de conceptos básicos (taxonomía) es usada en el contenido del mensaje para la acción cooperativa entre los agentes. Tomando como base la plantilla presentada por Moros(2013) para la descripción de estructura del lenguaje de comunicación entre los agentes, se muestra en el cuadro 22 la estructura del mensaje entre agentes de la arquitectura propuesta.

Cuadro 22
Estructura del mensaje entre agentes

Ítems	**Descripción**	**Ejemplo**
Tipo de mensaje	Este ítem responde a la pregunta: cuál es el propósito el mensaje?	De Información De Consulta
Actor/Role	Emisor o receptor, responde a la pregunta quien está manipulando el mensaje o role?	Agente director Agente Rastreador
Contenido	Representación del mensaje enriquecido semánticamente, precondiciones, post condiciones y acción.	RDF (Resource Description framework)
Lenguaje	Es la especificación de lenguaje de contenido de agentes ACL	FIPA-ACL-SL (lenguaje de comunicación de agentes semántico)
Protocolo de transporte	Especifica el servicio de transporte de mensajería de agentes donde utiliza el protocolo estándar MTP (Message transport protocol), especifica las diferentes redes de transporte en las cuales puede trabajar .	HTTP (Hypertext Transfer Protocol) IIOP (Internet InterORB Protocol) WAP (Wireless Application Protocol)

Fuente: El autor de la investigación

En la figura 24, se muestra el bosquejo general de todos los agentes que conforman la capa de adquisición de datos

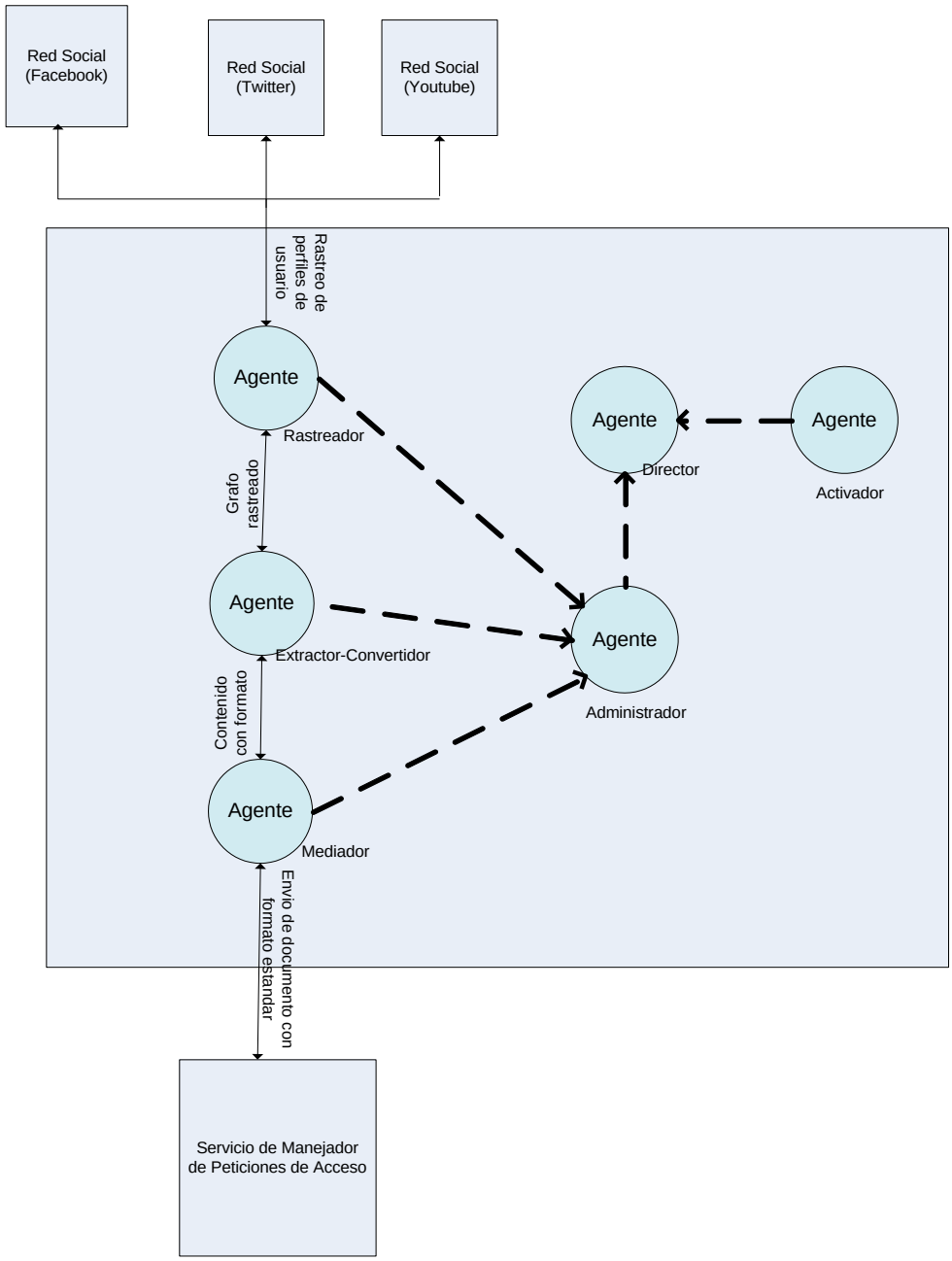


Figura 24. Agentes que intervienen en la Capa de Adquisición de Datos.
Fuente: El autor de la Investigación

Capa de Almacenamiento

La capa de almacenamiento va estar orientada a servicio, por ende para este apartado comenzamos con la selección de de los servicios candidatos.

De acuerdo Earl (2005), los procesos principales de un sistema servirán a priori como servicios candidatos de la arquitectura. A través de la caracterización de los trabajos de Boukhanovsky y Semenov (2011) y de Mysore y otros (2013) para trabajar con datos big data, se palpó (ver cuadro 23), los procesos importantes que intervienen en la alimentación, construcción y consulta de repositorio de datos generados en redes sociales.

Cuadro 23

Procesos en capa Almacenamiento

Procesos en Capa Almacenamiento	**Subprocesos**
Almacenar datos en repositorio no estructurado	-Activar la operación de agregar datos. -Instanciación del API de la BD para el acceso y manipulación de los datos. -Indexar repositorio. -Ejecutar el índice distribuido
Actualizar repositorio	-Activar la operación de actualizar datos. -Instanciación del API de la BD para el acceso y manipulación de los datos. -Indexar repositorio. -Ejecutar el índice distribuido
Manejar peticiones de acceso	-Procesar peticiones de nuevos documentos almacenar. -Gestionar cola de prioridades. -Procesar peticiones de consulta de datos. -procesar autorización de acceso

Procesos en Capa Almacenamiento	Subprocesos
Consultar información almacenada	- Determinar Relevancia - Ordenar Búsqueda - Mostrar Resultados
Recuperar datos almacenados	-Activar la operación de recuperar datos. -Instanciación del API de la BD para el acceso y manipulación de los datos. -Indexar repositorio. -Ejecutar el índice distribuido.
Ejecutar modelo MapReduce	-Ejecutar función Map -Obtener la data intermediaria -Ejecutar función Reduce

Fuente: Autor de la Investigación

Para tener una mejor percepción de los servicios que se mencionaron anteriormente, se agregó un nuevo proceso denominado Conector BD NoSQL, en donde se ejecuta los subprocesos comunes de comunicación con la base de datos, que son: instanciación del API de la base de dato NoSQL para la manipulación de los datos, indexar repositorio y ejecutar el índice distribuido, estos estarán acoplado bajo este nuevo proceso que tendrá la responsabilidad de encargase exclusivamente de la manipulación de los datos que se encuentren almacenados en la base de datos NoSQL. Es por tal motivo que los procesos anteriormente mencionados, fueron desacoplados de los procesos principales, que son: recuperar datos almacenados, actualizar repositorio y almacenar datos en repositorio no estructurado.

La identificación de una serie de servicios candidatos nos permitió de manera preliminar identificar las capas de servicio de negocio y de aplicación. Valera especifica (2013) que aquellos servicios que representan lógica genérica, reusable y neutral, pueden ser clasificados como servicios de aplicación, el conjunto de estos

establecen preliminarmente una capa de servicio de aplicación. Con respecto a los servicios de negocio, Valera (2013) comenta que los mismos contiene la lógica de flujo de trabajo usada para coordinar una composición de servicios subyacentes.

Valera(2013) continua que el rol principal de los servicios de negocio es actuar como controladores, componiéndose de servicios de aplicaciones para llevar a cabo la lógica de negocio requerida.

Para el caso de la solución propuesta, todos los servicios que conforman la capa de almacenamiento están conformado por servicios de aplicación, ya que cada servicio será accesado solo entre los servicios de la aplicación y cada uno de ellos corresponden a un proceso autónomo y desacoplado de la misma. En la figura 25 se muestra los servicios que conformara la capa de almacenamiento de la solución

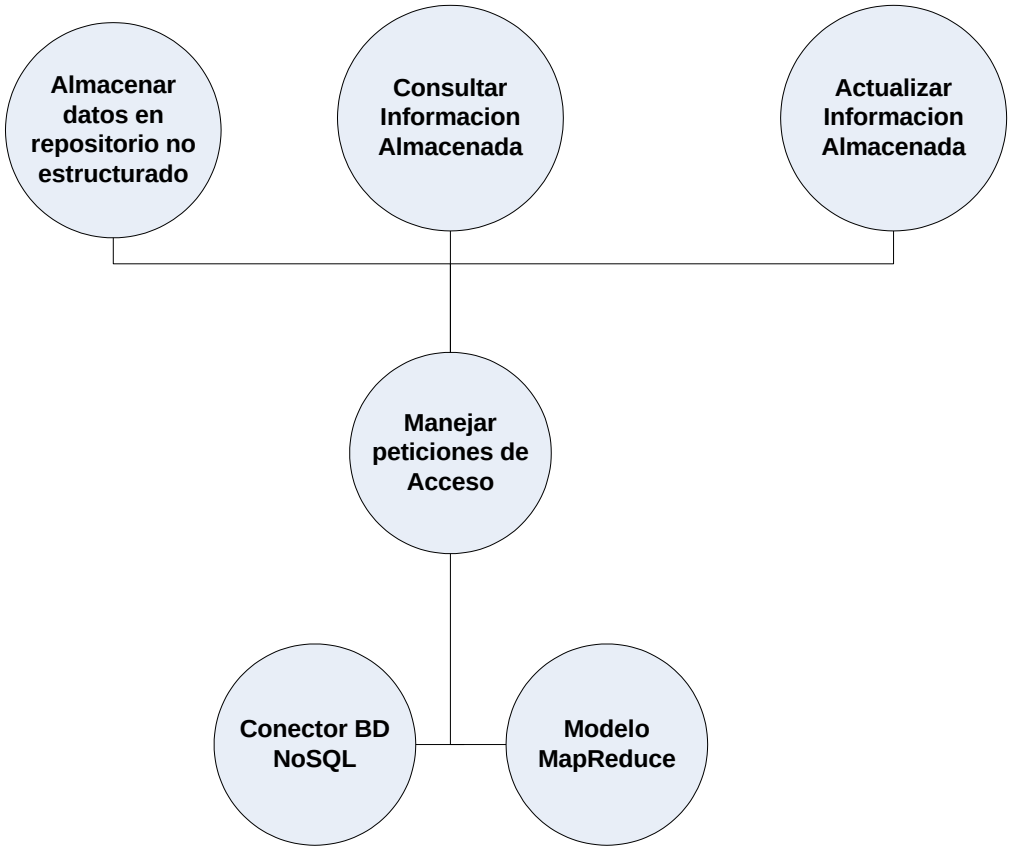


Figura 25. Servicios que intervienen en la Capa de Almacenamiento.
Fuente: El autor de la Investigación

Ya teniendo una descripción a nivel funcional de la solución propuesta, se procede a definir y diseñar la arquitectura a usar, el cual se llevará a cabo en el siguiente apartado. Pero primero presentamos los aportes de los estudios precedentes a esta investigación que ayudaron a definir la propuesta arquitectural.

Cuadro 24
Aportes de Estudio Anteriores

Antecedente	Aporte
Canali y otros(2011) *"Data Acquisition in Social Networks: Issues and Proposals"*	Identificación de los métodos de recuperación existente, como es el caso de la documentación del crawler, como técnica usada para obtener datos generados en redes sociales.
Catanese y otros(2011) *Rastreo de Facebook para fines de análisis de redes sociales*	Identificación de las acciones básicas que ejecutaran los componentes en este caso los agentes de rastreo y extracción de datos, que estarán habilitados en la arquitectura.
Boukhanovsky A , Semenov A, Veijalainen J. (2011) *Una arquitectura genérica para Monitorear y Analizar una Red Social*	Modelo Arquitectural y Modelo de requisito
Mysore, Khupat y Jain (2013) *Arquitectura Big Data y patrones, parte 4: Entendiendo patrones atómicos y compuestos para soluciones Big Data*	Plantilla arquitectural para modelar una solución big data. Componentes básicos que deben existir para conformar una arquitectura big data.
Mysore, Khupat y Jain (2013) *Arquitectura Big Data y patrones, parte 5: aplicar un patrón de solución a su problema Big Data y elegir productos para implementarlo*	Modelo Arquitectural Documentación de los procesos que intervienen en la gestión de datos no estructurados que se generan en redes sociales, estos son: captura, transformación y almacenamiento.
Morros (2013) *Big Data- Análisis de Herramientas y Soluciones*	Identificación de técnicas existentes, para tratar el almacenamiento de dato no estructurados
Souravlias et (2012) *InterSocialDB:Una infraestructura*	Identificación de los procesos mínimos que deben existir para la recolección y

para gestionar Datos sociales	almacenamiento de datos no estructurado generados en redes sociales, adicional se toma como base las combinaciones NoSQL descritas allí que se puede emplear para definir un repositorio de datos de tipo no estructurado.
Martínez (2013) *Desarrollo de una Herramienta de Inteligencia de Negocio para el análisis de redes sociales almacenada en grafos*	Se basa en la utilización de gestores de base de datos orientados a grafos, para el almacenamiento del grafo de la red social con sus respectivos nodos, con la finalidad de mantener el historial de los perfiles de usuarios ya visitados

Fuente: Autor de la investigación

Propuesta de la Arquitectura de Software

Uno de los ambientes de trabajo para representar arquitectura, construido sobre la base de UML y siguiendo el estándar IEEE 1471 (2000) es el propuesto por Garland (2003), el autor señala que su metodología desde sus comienzos fue pensada como una forma ágil de representar arquitectura, exponiendo que el arquitecto puede hacer uso de aquellos artefactos que verdaderamente presenten información útil al equipo de diseñadores.

Para definir la arquitectura de software correspondiente, se emplearán algunas de las vistas sugeridas por Garland (2003), en especifico aquellas que resulten de valor y relevancia para alcanzar los objetivos propuestos.

La figura 26, muestra un diagrama de contexto para entender que actores, procesos y sistemas externos, utilizan la arquitectura propuesta.

Unos de los pilares imprescindibles dentro de la metodología de Garland (2003), es la importancia del desarrollo basado en componentes para la construcción

de sistemas de software, por lo tanto es fundamental dar un panorama de todos los componentes que conformarán el sistema, como ellos se conectan y se despliegan en la infraestructura de hardware.

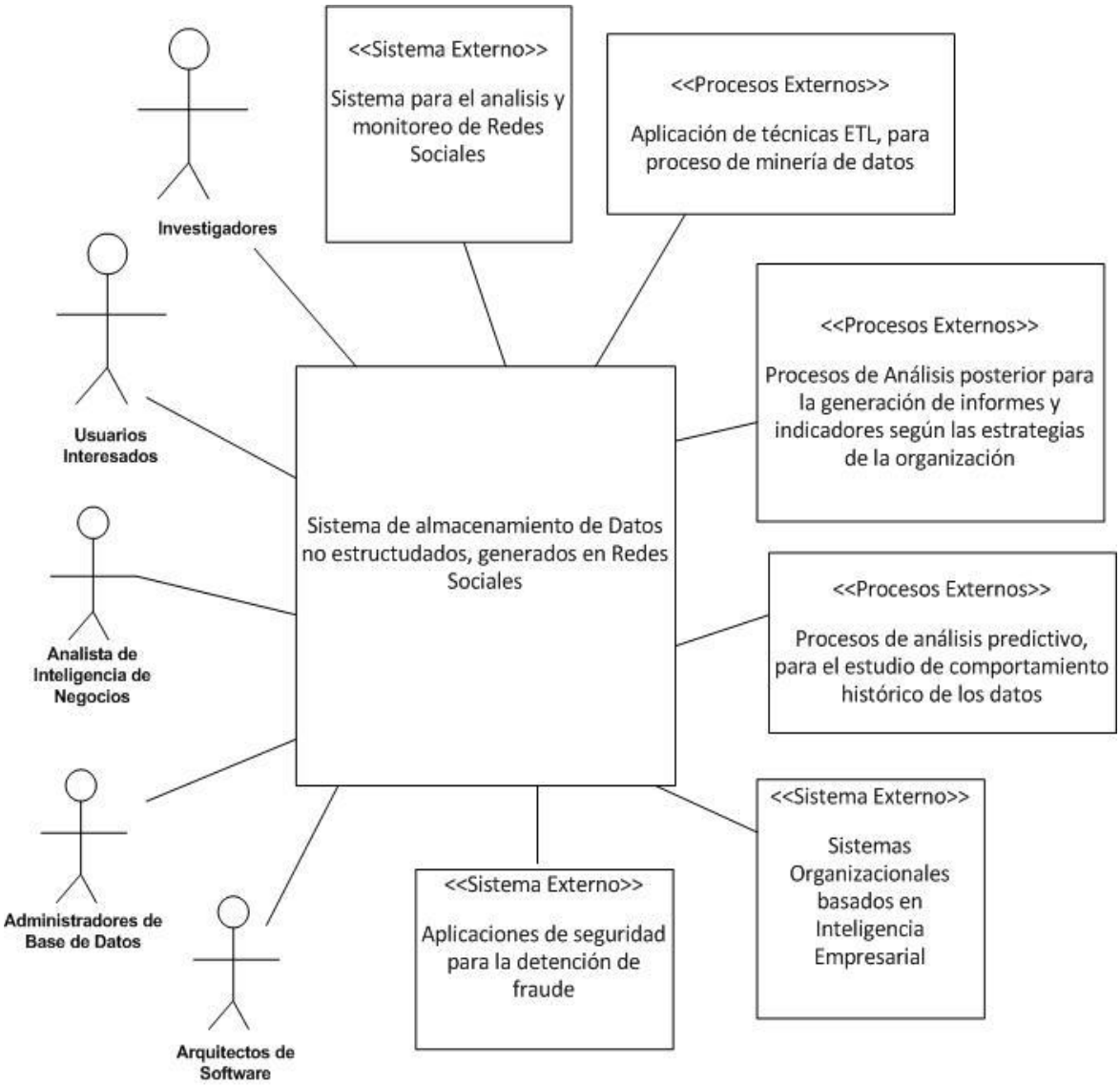


Figura 26. Diagrama de Contexto de la Arquitectura Propuesta.
Fuente: El autor de la Investigación

Garland(2003), establece que los componentes preliminares de la arquitectura pueden ser aquellos cuyas funcionalidades encapsula los casos de uso principales de la aplicación, asegurando que "La clave para la definición de la arquitectura exitosa usando componentes es que cada componente tiene configuraciones definidas y mecanismos de comunicación que pueden utilizarse para combinar un conjunto de componentes para lograr un conjunto de funciones del sistema."(p.114).

Por lo anterior comentado, se tomó el análisis de los casos de uso, con la finalidad de que cada uno de ellos servirán como componente de la aplicación. En función de reducir el espacio del nombre y colocarle un sujeto, fueron renombrados como se observa en el cuadro 25.

Cuadro 25
Componentes de la Arquitectura de Software

Nombre Caso de Uso	**Nombre de Componente**
Rastrear Información de Perfil Usuario en redes sociales	Rastreador de datos
Extraer Contenido generado en redes sociales	Extractor de datos
Convertir información en formato único	Convertidor de datos
Extraer Lista de amigos	Extractor de Relaciones
Almacenar Datos no Estructurados en el Repositorio	Almacenador de Datos
Actualizar Repositorio de Datos No estructurado	Actualizador Repositorio
Recuperar datos almacenados en el repositorio	Recuperador de datos
Manejar Peticiones de Acceso	Manejador de Peticiones
Consultar Contenido no estructurado	Buscador de Contenido

Almacenado	
Registrar Datos de usuario a rastrear	Registrar Datos
Autorizar rastreo de Perfil usuario	Autorizar Rastreo

Fuente: Autor de la Investigación

A continuación la figura 27, muestra el diagrama de componente de la arquitectura propuesta

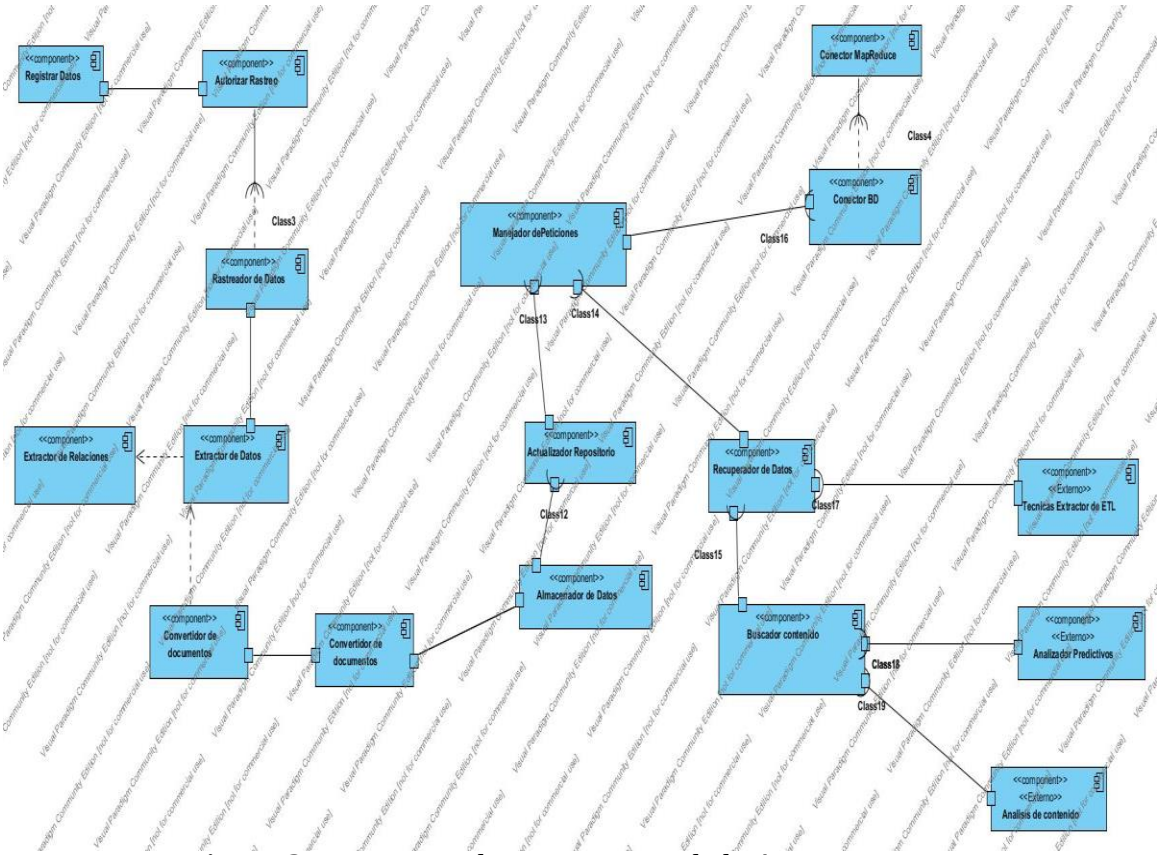


Figura 27. Diagrama de componentes de la Arquitectura Propuesta.
Fuente: El autor de la Investigación

En la figura 28, se muestra un bosquejo general de la solución arquitectural con todos los elementos descritos anteriormente.

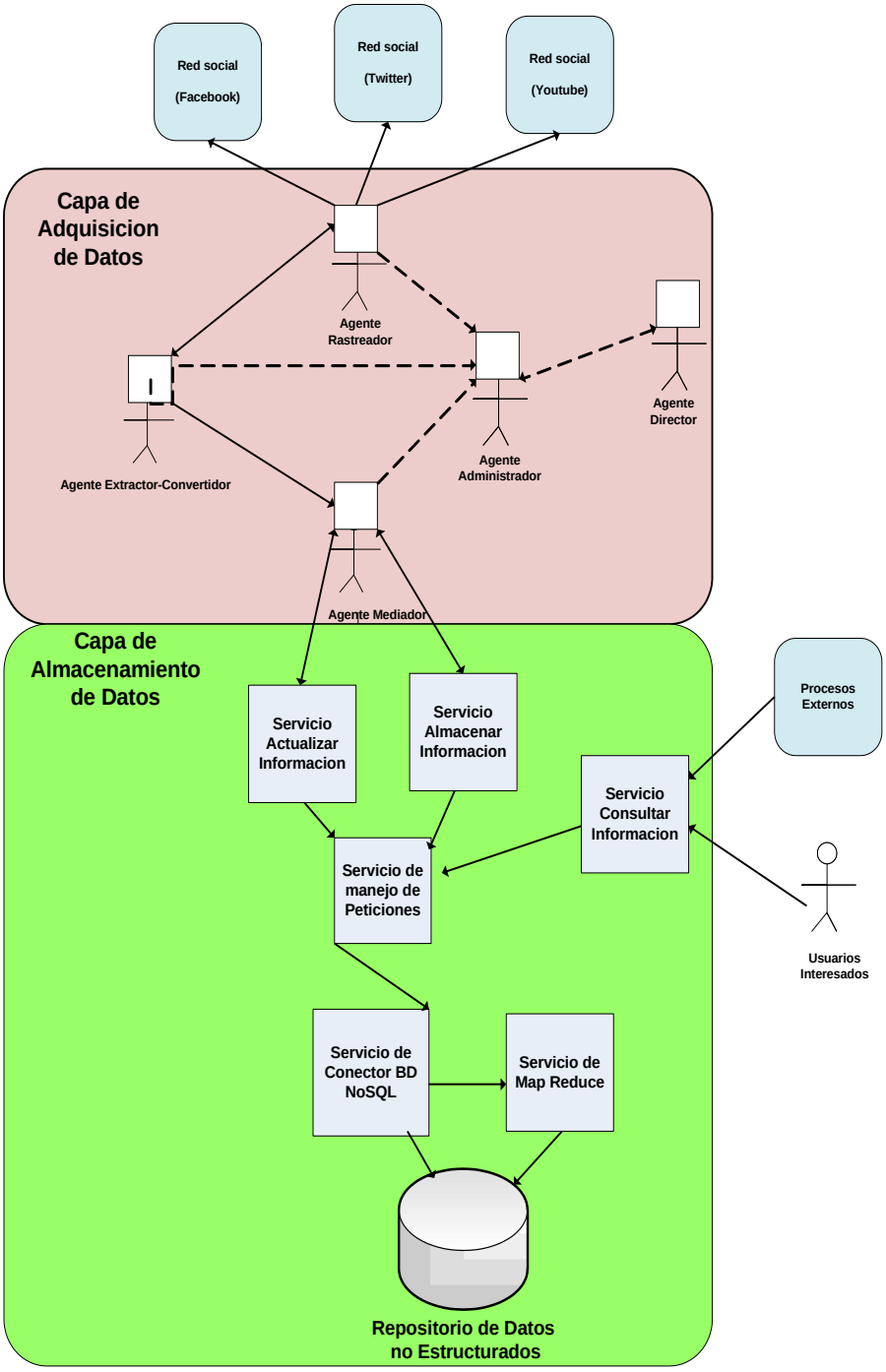


Figura 28. Diseño de la Arquitectura Propuesta.
Fuente: El autor de la Investigación

CAPITULO V

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

Una arquitectura de tipo big data, ofrece los beneficios de adquirir, procesar y almacenar los datos de tipo no estructurados de diversas formas. Cada fuente de big data se gestiona de distinta manera, en el ámbito de Redes sociales no es la excepción allí se observa que los datos generados allí presentan características de las tres V, volumen, velocidad y variedad, es por esta razón que se tomo el modelo big data como base para definir la arquitectura propuesta para la construcción de un repositorio de datos no estructurados generados en fuentes sociales. Es por tal motivo que se realizo una investigación bibliográfica en donde se estudio a fondo, los elementos que intervienen en una arquitectura Big data, la tecnología NoSQl, los métodos para modelar NoSQL, los diferentes contenido generados en redes sociales y los métodos de recuperación y búsqueda usados para los mismos. Así mismo para apoyar los procesos de captura y almacenamiento de la información gestionada se profundizo en los sistemas multiagente y en el diseño de servicios.

En este sentido, el presente estudio permite establecer las siguientes conclusiones:

1. El diseño de la arquitectura permitió identificar los principios fundamentales que intervienen en la construcción de un repositorio de datos generados en redes sociales,

los elementos esenciales que se deben considerar son : el rastreo de la red social en donde se genera el dato, la extracción del dato, la transformación del dato en un formato estándar, y finalmente el almacenamiento del mismo.

2. La capa de adquisición de datos presentada en la arquitectura es basada en multi-agentes, con el fin de proveer un procedimiento relevante de rastreo y captura de información que se genera en las distintas redes sociales.

3. Los agentes rastreadores están basado en el comportamiento de los crawler sociales, que en conjunto con las APIs sociales, demuestran que son los mecanismos de búsqueda y recuperación usados para los contenidos generados en redes sociales.

4. Cada uno de los aspectos desarrollados dio cumplimiento a los objetivos específicos establecidos en la investigación, partiendo del diagnóstico de los requerimientos funcionales y no funcionales para identificar cuáles son los elementos que intervienen en la construcción de un repositorio de datos enmarcado en redes sociales.

5.Basado en el principio de big data de almacenar grandes volúmenes de datos no estructurados en tiempo real, se diseño un repositorio ubicado en la capa de almacenamiento para almacenar los datos previamente extraído por los agentes rastreadores, con el fin de generar un historial de información que se genera en los distintos espacio sociales. El repositorio está fundamentado en tecnología NoSQL, el cual confirma que este tipo de tecnología es recomendable para gestionar almacenamiento de datos no estructurados que se genere en redes sociales.

6. El repositorio de datos no estructurados contemplado en la capa de almacenamiento, está conformado por dos base de datos, la primera es de tipo de documento en donde se almacena el contenido extraído y la segunda es de tipo grafo en donde se almacena datos del usuario en donde se ubico el contenido en la red

social. El fin de esto es mantener un historial de correlación de la información extraída y el autor (usuario)que la genera.

7. Como parte del diseño en la capa de almacenamiento se contemplo una serie de servicios los cuales se encargan de gestionar la persistencia de datos dentro de la arquitectura, esto es debido a que los contenidos extraídos son convertidos por el agente rastreador en un formato estándar de tipo documento para luego ser enviado a los servicios correspondientes para gestionar su respectivo almacenamiento.

8. El uso de técnicas de modelados para NoSQL mencionadas por Katsov y reforzadas por Fowler y Sadalage, permitió definir la estructura de la metadata a almacenar, garantizando así el acceso rápido a los datos en el repositorio, causando una ventaja para los procesos terceros de la solución, de reducir sus tiempos de consulta al momento de consultar los datos almacenados .

9. La metodología de diseño de arquitectura orientada a componentes, propuesta por Garland (2003), proporcionó una guía adecuada para entender cuáles eran los artefactos UML útiles para el desarrollo de la solución, tales como Diagramas de Casos de Uso, Clase, Componente y Despliegue, que facilitaron la percepción arquitectural del sistema.

10. El estudio, proporciono un espacio para dar un aporte acerca del comportamiento detallado que presenta el modelo Big data para cubrir soluciones que requieran tratar datos no estructurados, esta contribución se realizo a través de la descripción detallada de cada uno de los elementos que intervienen en el, permitiendo así generar una plantilla arquitectural especifica enfocada a la construcción de repositorios para datos no estructurados que se requieran utilizar en el ámbito de redes sociales y social media.

12. La construcción de un repositorio para datos no estructurados, generados en redes sociales, aporta al área de investigación relacionada al descubrimiento del conocimiento, una fuente integrada de datos con formato único, que se encuentran

listos para ser procesados por las tareas especificas de minería de datos. Tanto el repositorio como la minería de datos, representa una ayuda en este ámbito, ya que a través de ellos se realiza la búsqueda de patrones que se esconden en el enorme conjunto de datos, con la finalidad de interpretarlos y generar conocimiento para obtener información útil para la toma de decisión.

Recomendaciones

Con base a los planteamientos presentados anteriormente y que pretenden dar respuesta a los objetivos planteados en la investigación, es pertinente el planteamiento de algunas recomendaciones con las cuales se pueda mejorar la construcción de un repositorio de datos no estructurados generados en redes sociales:

1. Motivado a la naturaleza cambiante del dominio del problema planteado, y que los esfuerzos para obtener información generada en las redes sociales cada vez son más demandante, se debe continuar a identificar nuevos métodos o novedoso mecanismos de recuperación que pueden ir apareciendo aplicados para la extracción datos sociales.

2. Debido a que se están empezando a implementar métodos para modelar NoSQL para definir la metadata para repositorios que almacenen datos no estructurados, se ha creado un nuevo espacio en este campo de investigación que es el uso de patrones NoSQL, se sugiere investigar más sobre esto.

3. Estudiar más a fondo la integración de tecnologías NoSQL con otras tecnología como SOA, LAMP, entre otras, con el fin de mejorar la gestión de la persistencia de datos en la solución.

4. Debido a la demanda de integrar diversos datos generados no solamente de redes sociales sino de otras fuentes, se sugiere investigar más acerca de la persistencia Poliglota en Big Data.

5. Se sugiere de investigar un poco más a fondo acerca de la nuevos mecanismos de seguridad que van adoptando las redes sociales para la autorización del acceso de los lo información publicada por los usuarios participantes.

6. Mantener una constante actualización acerca de las nuevas tecnología NoSQL que puedan ir apareciendo para almacenar datos no estructurados.

7. Se sugiere profundizar un poco más en los mecanismos de gestión para tratar datos no estructurados (específicamente los de tipo audio y video) que se producen en tiempo real, como es el caso de los contenidos generados en plataforma de tipo streaming.

8. Se recomienda presentar el modelo arquitectural planteado usando una notación para modelar arquitecturas que nos ayude a mitigar los riesgo en función a los objetivos de atributo de calidad, ejemplo: ATAM (Architecture Trade-off Analysis Method).

REFERENCIAS BIBLIOGRÁFICAS

Cázeres, L., Christen, M., Jaramillo, E., Villaseñor, L. y Zamudio, L. E. (1980). Técnicas actuales de investigación documental. México: Universidad Autónoma Metropolitana.

Hyacinth S. Nwana (1996) Software Agents: An Overview . Knowledge Engineering Review, Vol. 11, No 3, pp.1-40, Sept 1996. © Cambridge University Press

Arias, F.(1999). El proyecto de investigación: guía de investigación. 3era Edición. Episteme.

Jiménez, W. (2000). Formulación de proyectos factibles para el área educativa. UPEL-IPB, Barquisimeto.

Felix, L. (2002). Minería de Datos. [Articulo en Línea] Disponible: http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html
[Consulta: Enero 21, 2014].

Universidad Centroccidental "Lisandro Alvarado" (UCLA). (2002). Manual para la Presentación del Trabajo Conducente al Grado Académico de: Especialización, Maestría, Doctorado. Barquisimeto. Venezuela.

Universidad Pedagógica Experimental Libertador (UPEL), Vicerrectorado de Investigación y Postgrado. (2002). Manual de Trabajo de Grado y Tesis Doctorales. Caracas: Fondo Editorial de la Universidad Pedagógica Experimental Libertador.

Garland, J., Anthony, R. (2003). Large-Scale Software Architecture. West Sussex, Inglaterra: Jhon Wiley and Son.

Rojas, M. y Garcia, J (2004). Introducción y Principios Básicos del Desarrollo de software basado en componentes. Pontificia Universidad Javeriana.Bogota. Colombia.

Rouse, M.(2005) Data Aggregation. [Articulo en Línea]. Disponible: http://searchsqlserver.techtarget.com/definition/data-aggregation. [Consulta: Diciembre 16, 2014].

Martinez, A. (2007). Contenidos Generados por Usuarios. Estrategia y Arquitectura de Red Cableuropa, S.A.U. (ONO)

Bhat, U., y Jadhav S. (2009). Moving Towards non-relational Databases. Articulo Usha Mittal Institute of Technology : p13.

Burgueño, P. (2009). Tipos y Clasificación de Redes Sociales. [Articulo en Línea] Disponibles:http://www.pabloburgueno.com/2009/03/clasificacion-de-redes-sociales/ [Consulta: Enero 21, 2013].

Codina, L, (2009). Ciencia 2.0: redes sociales y aplicaciones en línea para académicos. [Articulo en Línea]. Disponible: http://ddd.uab.cat/pub/artpub/2011/88755/hipertext_a2003n1a9/ciencia-2-0.html [Consulta: Septiembre 15, 2014].

Flores, J., Moran, J. y Rodriguez, J. (2009). Las Redes Sociales. [Articulo en Línea] Disponible: http://mc142.uib.es:8080/rid=1HY8TVCBB-15599LW-1S6Z/redes_sociales.pdf [Consulta: Septiembre 21, 2012].

Gutierrez, J., Penny, R., y Pérez, T. (2009). Redes sociales.[Presentación en Línea]. Disponible: http://www.slideshare.net/jorluguvi/redes-sociales-2388331 [Consulta: Septiembre 12, 2012].

Gasca, R, De la Rosa F., Ceballos F. (2009). Arquitectura de un Crawler para extraer recursos electrónico. Universidad de Sevilla. Sevilla. España.

Rojas, O., Mendoza, M., Martin M., Ponta M. (2009). Framework de evaluación de crawling focalizado distribuido. Universidad de Valparaíso. Valparaíso. Chile.

Castañeda, L. Castañeda, I (2010). Redes Sociales y otros tejidos online para conectar personas. [Documento en Línea].Disponible: http://mc142.uib.es:8080/rid=1MX54C554-WJ3R5J-2WQ/Redes_sociales%20y%20otros%20tejidos%20online.pdf .Universidad de Murcia. España [Consulta: Agosto 25, 2014].

Comision Europea (2010). Social Networks Overview:Current Trends and Research Challenges. Information Society and Media

Ghaderi M, Yazdani, N. y Moshiri, B. (2010). A Social Network-based Meta Search Engine. Universidad de Teheran. Teheran. Iran.

Kaplan, A. y Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, Vol. 53, Issue 1, p. 59-68.

Guervos J. Esparcia A (2010). Usando bases de datos NoSQL para algoritmos evolutivos- p7.

Mújica, J. (2010). Redes sociales: historia, oportunidades y retos.[Articulo en Línea].Disponible:http://www.forumlibertas.com/frontend/forumlibertas/noticia.php?id_noticia=16428 [Consulta: Abril 23, 2013].

NeoHumano (2010). Conociendo el Social Media Monitoring. [Documento en Línea].Disponible:http://www.slideshare.net/neoconsulting/social-media optimizationmonitoring/ [Consulta: Octubre 30, 2012].

Rodriquez J (2010). Introducción a la base de datos NoSQL MongoDB: Instalación, primeros pasos y ejemplo de conexión con Java. [Documento en Línea]. Disponible:http://ubuntulife.wordpress.com/2010/04/13/introduccion-a-la-base-de-datos-nosql-mongodb-instalacion-primeros-pasos-y-ejemplo-de-conexion-con-java/ [Consulta: Febrero 12, 2012]

Steffens, H. (2010). Una breve historia de Social Media. [Documento en Línea]. Disponible:http://pulsosocial.com/2010/11/26/una-breve-historia-de-social-media/ [Consulta: Marzo 21, 2013]

Valera, R. (2010). Arquitectura de Software para Automatizar los registros Académicos en la Universidad Centroccidental "Lisandro Alvarado". Trabajo de ascenso, Universidad Centroccidental "Lisandro Alvarado", Barquisimeto, Venezuela.

Boukhanovsky A., Semenov A. y Veijalainen J (2011). "A generic Architecture for a Social Network Monitoring and Analysis System". International Conference on Network-Based Information Systems. The National Research University of Information Technologies, Mechanics and Optics (University ITMO). San Peterbusgo, Rusia.

Canali, C., Colajanni, M. y Lancellotti R. (2011). Data Acquisition in Social Networks: Issues and Proposals. Department of Information Engineering. University of Modena and Reggio Emilia. Italia.

Catanese, S., De Meo, P., Ferrara, E. y Fiumara, G. (2011). Crawling Facebook for Social Network Analysis Purposes. University of Messina, Italy.

Gomez, E. (2011). Modelo Arquitectural para Aplicaciones Móviles usando el enfoque de Líneas de Producción Dinámica de Software. Trabajo de Grado de Maestría. Universidad Centroccidental Lisandro Alvarado, Barquisimeto, Venezuela.

Iragorri, C.(Entrevistador). (2011, Octubre 21). Club de Prensa de NTN24: Entrevista a Antonieta Cádiz y Sagrario Ruiz de Apodaca Análisis de las redes sociales como fuentes de información [Video en línea]. Washington: Canal Nuestra TeleNoticias NTN24.Disponible: http://www.youtube.com/watch?v=Y1NA7MuF23E [Consulta: 2012, Febrero 22].

Kaushik, A. (2011). Best Social Media Metrics: Conversation, Amplification, Applause, Economic Value. [Articulo en Línea]. Disponible: http://www.kaushik.net/avinash/best-social-media-metrics-conversation-amplification-applause-economic-value/ [Consulta: Marzo 08 , 2013]

Muñoz, G. (2011). Mide y Analiza: medir en Redes Sociales [Blog en Línea]. Disponible: http://www.territoriocreativo.es/etc/2011/03/mide-y-analiza-medir-en-redes-sociales.html [Consulta: Marzo 08 , 2013]

Rooter. (2011). El Futuro del Derecho Autor y los Contenidos generados por los usuarios en la web 2.0. [Documento en Línea]. Disponible: http://rooter.es/documents/futuro_derechos_autor_contenidos_generados_usuarios_web_2.0.pdf. [Consulta: Noviembre 28, 2013].

Ruiz, V (2011). Apis de Medios Sociales. [Documento en Línea]. Disponible: http://es.slideshare.net/rvr/apis-de-medios-sociales. [Consulta: Agosto 15 , 2013].

Valera, E. (2011). Construcción de un motor de Búsqueda de Contenidos en repositorios confiables, basado en crawlers, enmarcado en una arquitectura orientada a servicio (SOA). Trabajo de Grado de Maestría. Universidad Centroccidental Lisandro Alvarado, Barquisimeto, Venezuela.

Villar, A. (2011). 5 metricas para medir tus esfuerzos en Redes Social. [Blog en Línea]. Disponible: http://abrahamvillar.es/2012/06/5-metricas-para-medir-tus-esfuerzos-en-social-media/ [Consulta: Marzo 12, 2013].

Chaudhuri, S. (2012). What Next? A Half-Dozen Data Management Research Goals for Big Data and the Cloud. Microsoft.E.U.U.

Gundecha, P., Liu, H (2012). Mining Social Media: A Brief Introduction.Universidad de Arizona. E.U.U.

Cuervo, M. Sanabria, J. y Romero, A. (2012). Utilidad y funcionamiento de las bases de datos NoSQL. Revista Facultad de Ingeniería, vol 21, No33. Universidad Pedagógica y Tecnológica de Colombia. Bogotá, Colombia.

Katsov, I (2012) . NoSQL Data Modeling Techniques. [Articulo en Línea]. Disponible: https://highlyscalable.wordpress.com/2012/03/01/nosql-data-modeling-techniques/ [Consulta: Septiembre 12, 2013].

Fragoso, R (2012). ¿Qué es Big Data?. [Articulo en Línea].Disponible: http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/.[Consulta: Diciembre 12, 2013].

Koloniari, G. , Souravlias, D. y Pitoura, E. (2012). InterSocialDB: An Infrastructure for Managing Social Data. DMOD Laboratory. University of Ioannina. E.U.U.

Paulo, R. Costa, S., Souza, F. y Benevenuto, F (2012). Towards Integrating Online Social networks and Business Intelligence. Universidad de Federal de Pernambuco. Recife. Brasil.

Polisience (2012). Repositorios Definicion. [Blog en Línea]. Disponible: http://poliscience.blogs.upv.es/open-access/repositorios/definicion-y-tipos/[Consulta: Abril 12, 2014].

Sun, H., Heller, P. (2012) Oracle Information Architecture:An Architect's Guide to Big Data. [Articulo en Línea]. Disponible: http://www.oracle.com/technetwork/topics/entarch/

/articles/oea-big-data-guide-1522052.pdf

Stavrakantonaki, J. y otros. (2012). An approach for evaluation of social media monitoring tolos. 1st International Workshop on Common Value Management CVM2012. Heraklion**.** Grecia. pp. 52 -61.

Adedoyin, M., Medhat M, Stahl F (2013). A Survey of Data Mining Techniques for Social Network Analysis. School of Computing Science and Digital Media. Ucrania

Amador, G. (2013). Venezuela tiene más 12 millones de usuarios de internet. [Articulo en Línea]. Disponible: http://www.lanacion.com.ve/tecnologia/venezuela-tiene-mas-12-millones-de-usuarios-de-internet/ [Consulta: Mayo 28, 2013].

Bhagubhai, K. (2013). Use Big Data Technologies to Build a Content Repository Architecture. [Articulo en Línea] . Disponibles : http://www.devx.com/enterprise/big-data-technologies-content-repository.html. [Consulta: Junio 01, 2014].

Chulis, K. (2013). Desarrollo de datamarts de medios sociales utilizando herramientas de minería de textos de SPSS. [Articulo en Línea]. . Disponibles: http://www.ibm.com/developerworks/ssa/library/ba-social-media-spss-text-mining/ [Consulta: Noviembre 01, 2014].

ComScore. (2013). Futuro Digital Latinoamérica 2013. [Documento en Línea]. Disponible:http://www.comscore.com/lat/Insights/Blog/2013_Digital_Future_in_Focus_Series/ [Consulta: Mayo 25, 2013].

Díaz, W. (2013). Bases de Datos NoSql: llegaron para quedarse. [Documento en Línea]. Disponible: http://basesdedatosnosql.blogspot.com/ [Consulta: Agosto 8, 2013].

Fowler, M y Sadalage. P (2013). NoSQL Distilled : A Brief Guide to the Emerging World of Polyglot Persistence.vol 1, No 335, Editorial : Addison-Wesley.

IRedes. (2013, Marzo). Presentación de la tercera versión del Mapa iRedes [Transcripción en Línea]. Ponencia presentada en III Congreso Iberoamericano sobre Redes Sociales. Disponible: http://www.iredes.es/2013/03/tercera-version-del-mapa-iredes/ [Consulta: Mayo 23, 2013].

Garcia, E. (2013). Concepto Big Data. [Articulo en Línea]. Disponible:http://www.cnis.es/index.php?option=com_content&view=article&id=434:concepto-big-data&catid=47:noticias-boletin-cnis&Itemid=57[Consulta: Enero 23, 2014].

Gracia, L. Eligiendo una Base de Datos NoSQL según el Teorema CAP. [Articulo en Línea]. Disponible: https://unpocodejava.wordpress.com/2013/05/29/eligiendo-una-base-de-datos-nosql-segun-el-teorema-cap/ [Consulta: Octubre 15, 2014].

Martin, L. (2013). Principales tecnologías en Big Data:NoSQL. [Articulo en Línea]. Disponible:http://www.brainsins.com/es/blog/principales-tecnologias-big-data-nosql/107943 [Consulta: Octubre 23, 2014].

Michael, K., Miller, K.(2013). Big Data: New Opportunities and New Challenges. Universidad de Wollongong y Universidad de Missouri–St. Louis.E.U.U

Martínez, N. (2013). Desarrollo de una Herramienta Business Intelligence para el análisis de redes sociales almacenadas en grafos. Universidad de Cantabria. España.

Mercado, J. (2013). Too Big to Ignore: BIG DATA. Artículo Científico en Línea. Disponible: http://deloitte.wsj.com/cfo/files/2013/09/TooBigIgnore.pdf´. [Consulta: Diciembre 01, 2013].

Morros, R. (2013). Big Data- Análisis de Herramientas y Soluciones. Proyecto de Final de Carrera. Everis – Faculta de Informática de Barcelona – UPC. Barcelona. España.

Mysore, D., Khupat, K., Jain, S.(2013).Big data architecture and patterns, Part 4:Understanding atomic and composite patterns for big data solutions. IBM developerWorks.[Artículo en Línea]. Disponible: http://www.ibm.com/developerworks/ library/bd-archpatterns4/bd-archpatterns4-pdf.pdf. [Consulta: Mayo 01, 2014].

Mysore, D., Khupat, K., Jain, S.(2013).Big data architecture and patterns, Part 5: Apply asolution pattern to your big data problem and choosethe products to implement it. [Artículo en Línea]. Disponible: http://www.ibm.com/developerworks/library/bd-archpatterns5/bd-archpatterns5-pdf.pdf. [Consulta: Mayo 01, 2014].

Sandoval, J. y Jiménez F. (2014). APIs de Redes Sociales. [Artículo en Línea]. Disponible: http://sg.com.mx/revista/45/apis-redes-sociales#.VDw7OvmSxSA. [Consulta: Diciembre 01, 2014].

ANEXOS

A. Currículum Vitae del Autor

Maria Esperanza Linarez, portador de la C.I: N° V-17.157.351, nace en Barquisimeto el 8 de Diciembre de 1984. Realiza estudios de primaria en la Unidad Educativa Colegio "Valle de Cabudare". Obtiene el título de bachiller en el Colegio "Francisco Tamayo . Ingresa a la universidad Centroccidental "Lisandro Alvarado" en el año 2003, para iniciar estudios de pregrado en la carrera de Ingeniería en Informática, egresando en el año 2008 y obteniendo el décimo tercero (13ro) lugar de la promoción. Seguidamente en el año 2009 realiza un diplomado en el área de base de datos, para reforzar y aprender un poco más acerca de tareas de mantenimiento y administración en sistema de gestión en base de datos. En el año 2009 comienza su actividad profesional empezando como técnico de soporte en la Gobernación del Estado Lara. Luego de un año, pasa en el 2010, a ejercer la coordinación de la unidad de Informática en el Servicio Desconcentrado Oncológico del Estado Lara (SAO) rol que cumple hasta mediados del 2012. En ese tiempo se traslada a la ciudad de valencia, en donde ingresa como desarrollador de aplicaciones en la empresa Vamatech C.A, empresa especializada en ofrecer servicios informáticos de alta tecnología a diversos sectores empresariales. Luego de cumplir un año en dicha empresa se traslada nuevamente a Barquisimeto en donde labora hoy en día como especialista de aplicaciones, en una empresa en el área de tecnología reconocida por prestar servicios de tecnología de información para procesos de negocio. Durante su carrera de pregrado obtuvo varias distinciones de cuadro de honor por rendimiento académico; y en año 2011 a nivel profesional obtuvo un reconocimiento por parte de la presidencia del Servicio Desconcentrado Oncológico del Estado Lara (SAO) por su buena disposición y dedicación a contribuir con el servicio.