Anexo 3: Estado del arte

Ing. Wilson Alzate Calderón[*1] and Ing. Alexandra Pomares Ph.D.[**1]

[1]Departamento de Ingeniería de Sistemas, Pontificia Universidad Javeriana

5 de diciembre de 2014

Índice

[*]walzate@javeriana.edu.co
[**]pomares@javeriana.edu.co

Índice de figuras

Índice de cuadros

1. BigData

En la actualidad, el monto de información almacenada en el mundo se estima alrededor de 1.200 exabytes(10^{18} bytes) [18], los cuales son producidos por múltiples fuentes; por ejemplo, Google procesa más de 24 petabytes(10^{15} bytes) de datos por día y Facebook obtiene más de 10 millones de fotos cada hora [18]. En el mundo de los negocios el 80 % de los datos existen en un formato no estructurado [23] y si este monto fuera solamente libros impresos, ellos cubrirían la superficie entera de Estados Unidos en 52 capas de grueso [18]. Se estima que en algún momento entre 2003 y 2004, el monto de datos creados por el mundo digital aceleró exponencialmente, sobrepasando el monto total de datos creado en los 40.000 años previos de la civilización humana [29].

Teniendo en cuenta estas grades cantidades de datos que se están generando actualmente, surge el paradigma de Big Data, que es un concepto generalmente usado para describir montos vastos de datos diversos, tanto estructurados como no estructurados, a los cuales las organizaciones pueden acceder de manera rápida, para analizarlos usando herramientas innovadoras, que en conjunto, ayudan a determinar con precisión oportunidades de mejora en la gestión y generación de valor [22].Un problema de Big Data es aquel que cumple con las siguientes tres características[19]:

- Volumen (Escala de los datos)
  Big Data trabaja con grandes cantidades de datos, lo que en la actualidad es algo frecuente teniendo en cuenta que por ejemplo para 2012 cerca de 2.5 exabytes de datos fueron creados cada día y ese número se está duplicando cada 40 meses [19].

- Variedad (Diferentes formas de datos)
  Existen diferentes formas de datos, tanto estructurados como no estructurados, dentro de los que se encuentran: mensajes, actualizaciones, imágenes en redes sociales, lecturas de sensores, señales de GPS desde teléfonos celulares, compras en línea, entre otros [19].

- Velocidad (Análisis de flujo de datos)
  Existen diferentes formas de datos, tanto estructurados como no estructurados, dentro de los que se encuentran: mensajes, actualizaciones, imágenes en redes sociales, lecturas de sensores, señales de GPS desde teléfonos celulares, compras en línea, entre otros [19].

El cuidado de la salud, en común con muchas otras industrias, está generando grandes cantidades de datos rutinarios, los cuales pueden ser minados

e incluso combinados con tweets y blogs. Es un gran reto procesar, analizar y conservar dicha masa de datos. Darle sentido a esa gran cantidad de información ofrece oportunidades para el mejor tratamiento de una enfermedad, abordar temas de salud pública o para el funcionamiento eficiente de los proveedores de servicios de salud [17]. El análisis de datos contenidos en las historias clínicas electrónicas usando técnicas computacionales es un problema de Big Data [23] debido al gran volumen de información contenida, a la variedad entre datos estructurados y no estructurados, así como también, la velocidad con la que se requieren ciertos análisis. A continuación se presenta un conjunto de aplicaciones existentes de Big Data en el dominio del cuidado de la salud.

1.1. Big Data en Salud

Luego de realizar una revisión bibliogáfica en las bases de datos ISI Web of Science y Scopus usando los siguientes criterios:

- ISI Web of Sciente
  TI=("big data" and "health") or TS=("big data" and "health")
- Scopus
  TITLE-ABS-KEY("big data" AND health) AND PUBYEAR >1999 AND (LIMIT-TO(DOCTYPE, "ar"))

se obtuvo un total de cien artículos científicos a los cuales se les efectuó una revisión por título, llegando a cincuenta ejemplares que fueron clasificados en cinco categorías: aplicaciones, conceptos, dilemas éticos, herramientas y tendencias. A continuación se presentan los aportes más relevantes:

1.1.1. Uso de Big Data y Cloud Computing para BI operacional en Salud

Se propone el uso de Cloud Computing para proveer computación como un servicio utilitario que puede servir de puente para superar la brecha digital para los Sistemas de Información Médica en los países en desarrollo. Se plantea cómo ciertas herramientas pueden explotar las capacidades de la computación en la nube para realizar análisis en Big Data. Usan un Health Management Information System (HMIS) llamado DHIS2 [1] y se analiza

[1] Es una herramienta para la recolección, validación, análisis y presentación de datos estadísticos adaptado (pero no limitado) a actividades de la gestión de información médica integrada

cómo sus módulos de Inteligencia de Negocios operacional son usados en los países en vía de desarrollo para gestionar sistemas a nivel de un país o un estado [27].

En los países en vía de desarrollo, donde se tienen sistemas EMR (Electronic Medical Record) pequeños, el manejo de Big Data es poco viable en términos de costos, por lo cual se propone el uso de AaaS (Analytics as a Service) con el fin de poder manejar grandes cantidades de transacciones de una terminología médica que no es estándar en registros médicos con una inversión razonable. Proponen además implementar terminologías médicas compartidas entre proveedores, para poder así hacer frente al reto de la variedad y con los modelos de nube AaaS manejar fácilmente los problemas de volumen y velocidad [27].

1.1.2. Big Data en cuidado de la salud personalizado

Este proyecto propone un *framework* llamado Collaborative Assessment and Recomedation Engine (CARE) que hace uso de Big Data en un modelo centrado en el paciente para crear un perfil de riesgo de enfermedades personalizado, así como también, un plan de gestión y de bienestar para un individuo usando una técnica de minería de datos llamada filtrado colaborativo. El objetivo de esta técnica es predecir la opinión de un usuario acerca de un ítem o servicio basado en las preferencias conocidas de un gran grupo de usuarios, que es la lógica en las recomendaciones de películas en Netflix.com o de libros en Amazon.com. Para ello se revisan las similitudes en estilo de vida, factores ambientales y predisposiciones genéticas que hacen que ciertas personas tengan mayor probabilidad de padecer enfermedades similares [5].

1.1.3. Big Data en Ecocardiografía Funcional

En este proyecto se hace uso de Ecocardiografía Funcional[2] para el apoyo en el análisis y diagnóstico de enfermedades. Se tiene como particularidad el uso de tecnologías de Big Data, lo cual fué necesario debido a que se decidió ampliar el número de variables a tener en cuenta para el análisis, ya que a menudo es el grupo de factores de deformación, en lugar de uno solo en particular, lo que identifica la presencia fenotípica de una enfermedad [29].

Para ello se propone el uso de las siguientes tecnologías:

1. Eficiencia de datos usando automatización basada en la nube: Uso de almacenamiento en la nube para la gran cantidad de datos obtenidos

[2]También conocida como ultrasonido cardíaco o electrocardiograma

2. Uso de robots y automatización en ultrasonidos cardiacos: Mayor eficiencia y eficacia en la toma de exámenes médicos

3. Computación cognitiva: Uso de técnicas de IA

4. Wearable computers: Dispositivos que se pueden llevar como ropa y la forma en que ellos ayudarán a obtener aún más información

1.1.4. Adquisición, compresión, encripción y almacenamiento de Big Data en Salud

Se propone un formato de compresión (Multiscale Electrophysiology Format) para la gran cantidad de datos obtenidos en registros electrofisiológicos[3] combinado con un sistema de adquisición de información y una base de datos de gran escala SAN (Storage Area Network).

El formato MEF cumple con la HIPAA (Health Insurance Portability and Accountability Act) en términos de protección de la información de cualquier paciente transmitida sobre una red pública. Se habla de la ventaja de usar este formato, ya que gracias a él, los investigadores pueden contar con todos los datos sin tener que eliminar variables por temas de almacenamiento, ya que los archivos quedan en un 20 % de su tamaño original [3].

1.1.5. Big Data en dispositivos médicos

Se utilizan sensores para recolectar información del funcionamiento de los aparatos médicos con el fin de obtener mayor confiabilidad y disponibilidad a bajo costo. Usaron tecnologías de Big Data para predecir el tiempo de vida útil restante, prevenir fallos en el servicio, mantenimientos no planeados y especialmente fallos que pudiesen causar serias pérdidas humanas y económicas [20].

1.1.6. Desarrollo de un ecosistema de cuidado de la salud usando IPHIs

Se propone el desarrollo de un ecosistema de comunicación de Big Data usando Procesadores Intermedios de Información de Salud (IPHI por sus siglas en inglés) que soporten el uso de fuentes de datos heterogéneas. El objetivo es pasar de un modelo de interoperabilidad ineficiente como el de la figura 1 a uno como el de la figura 2. Dichos IPHIs pueden también tener

[3]Electrofisiología: Es el estudio de las propiedades eléctricas de las células y tejidos biológicos.

observaciones de pacientes, comentarios de los medios sociales y sus salidas pueden también ir a la prensa o al público. Un ecosistema de información es un ambiente complejo en el cual los datos, los proveedores de información, los usuarios y los procesadores interactúan en un proceso mutuamente interdependiente y transformacional [17].

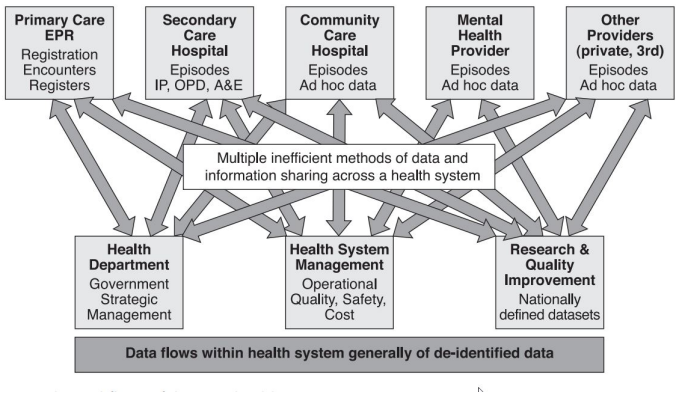

Figura 1: Flujos tradicionales de datos en un sistema de salud

Los IPHI se encontrarán entre los generadores de la información médica, a menudo los proveedores de servicios de salud y los usuarios de dichos datos. Los usuarios son los gerentes de servicios de salud, comisionistas, generadores de políticas, investigadores, farmacéuticas y otras industrias relacionadas a la salud [17], como se muestra en la figura 2

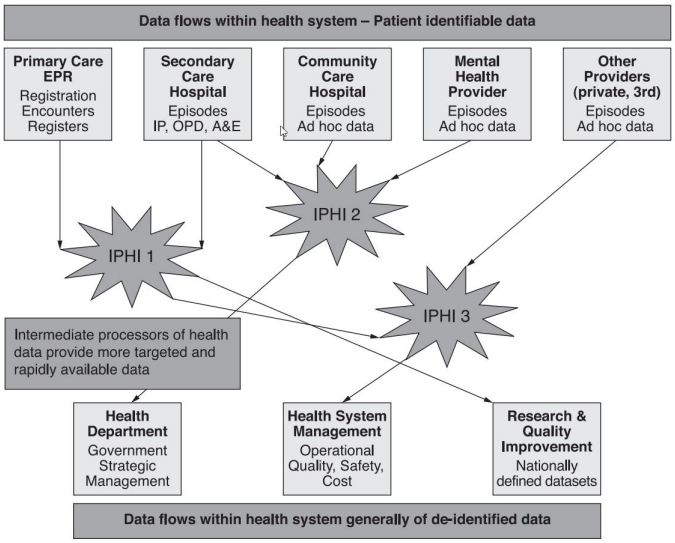

Figura 2: Rol de los IPHI en el sistema de salud

La implantación de dichos IPHIs en la industria de la salud, a diferencia de otros sectores, necesita hacer prevalecer la privacidad, asegurando a los profesionales, pacientes y el público en general, las disposiciones adecuadas para la gobernanza y seguridad de la información. Un ecosistema de salud de este tipo maximizaría el uso de los datos, creando nuevo conocimiento [17].

1.1.7. Big Data para evaluar el impacto de programas médicos

Se propone la combinación de datos públicos relacionados a un individuo con información vinculada a su condición médica, con el fin de entender de una mejor manera cómo destinar y retener pacientes reales, apalancándose en dichos datos para construir modelos predictivos que pueden asistir de una manera más adecuada a los médicos y realmente impactar en el comportamiento de pacientes con enfermedades crónicas [8]. Dentro de las cosas que se podrían evaluar se tiene:

- Identificar a individuos que tienen mayor probabilidad de ser beneficiados por un tipo particular de programa.
- Identificar individuos que tengan mayor probabilidad de participar activamente en un programa de manejo de enfermedades y qué nivel de divulgación será requerido para asegurar su participación.

- Identificar las razones que impactan más en la adherencia y cumplimiento de un paciente
- Identificar los tipos específicos de divulgación y apoyo que pueden tener mayor impacto en el comportamiento de un individuo y sus resultados.

Finalmente en el cuadro 1 se realiza una comparación de los artículos antes mencionados, encontrando que en su mayoría son propuestas teóricas de usos deseables de las tecnologías de Big Data en el dominio del cuidado de la salud. De igual manera son pocas las aproximaciones prácticas y palpables del uso de este paradigma en productos o prototipos que realmente hagan realidad los beneficios esperados. También se pudo evidenciar que uno de los factores recurrentes en los artículos, y por lo cual se habla de la necesidad del uso del paradigma Big Data en salud, es la alta cantidad de datos que se requiere analizar para que el personal médico y administrativo pueda tomar las mejores decisiones, entendiendo que una de las mayores fuentes de información es el texto narrativo consignado en las historias clínicas electrónicas. Por lo anterior es importante realizar una revisión de los procesos que pueden ser soportados (Capítulo 2), las tecnologías (Sección 1.2) y herramientas (Sección 1.3) asociadas al paradigma que pueden ser usadas.

Artículo	Tecnología	Problemática	Tipo	Contenido
Uso de Big Data y Cloud Computing para BI operacional en Salud (Sección 1.1.1)	Cloud Computing	Superar la brecha digital para los Sistemas de Información Médica en los países en desarrollo	Práctico	Texto
Big Data en cuidado de la salud personalizado (Sección 1.1.2)	Minería de datos	Perfil de riesgo de enfermedades personalizado	Teórico-Práctico	Texto
Big Data en Ecocardiografía Funcional (Sección 1.1.3)	Cloud computing, robots, Inteligencia Artificial	Poder contar con más variables en la ecocardiografía funcional	Teórico	Imágenes
Adquisición, compresión, encripción y almacenamiento de Big Data en Salud (Sección 1.1.4)	Procesamiento de archivos	Compresión de datos obtenidos en registros electrofisiológicos	Práctico	Imágenes
Big Data en dispositivos médicos (Sección 1.1.5)	Sensores	Confiabilidad y disponibilidad de dispositivos médicos	Teórico	Texto
Desarrollo de un ecosistema de cuidado de la salud usando IPHIs (Sección 1.1.6)	Ontologías	Ecosistema de interoperabilidad en salud para sistemas heterogéneos	Teórico	Texto
Big Data para evaluar el impacto de programas médicos (Sección 1.1.7)	Analytics	Destinar y retener pacientes reales	Teórico	Texto

Cuadro 1: Comparación de artículos de Big Data en salud

1.2. Tecnologías relacionadas a BigData

Como se mencionó anteriormente, cualquier problema de Big Data tiene tres características fundamentales como lo son la velocidad, la variedad y el volumen. Pero dado que Big Data no es como tal una tecnología, sino más bien un paradigma que debe ser atacado por medio de un conjunto de soluciones, se presentan a continuación algunas de las aproximaciones que pueden ayudar a abordar las propiedades de un problema de este tipo:

Tecnología	Sub-tecnología	Velocidad	Variedad	Volumen
NoSQL			X	X
Cloud computing	Infraestructure as a Service (IaaS)	X		X
	Software as a Service (SaaS)	X		X
	Platform as a Service (PaaS)	X		X
	Analytics as a Service (AaaS)	X	X	X
Analytics	Exploratory Data Analysis (EDA)		X	X
	Confirmatory Data Analysis (CDA)		X	X
	Qualitative Data Analysis (QDA)	X	X	X
MapReduce		X		X
Apache Hive		X		X
Apache Pig		X		X

Cuadro 2: Clasificación de tecnologías asociadas a Big Data

1.2.1. Cloud Computing

Es un modelo que permite acceso de red bajo demanda de manera conveniente y ubicua a un conjunto compartido de recursos de cómputo configurables (por ejemplo: redes, servidores, almacenamiento, aplicaciones y servicios) que pueden ser rápidamente aprovisionados y liberados con mínimo esfuerzo administrativo o interacción de un proveedor de servicios [27].

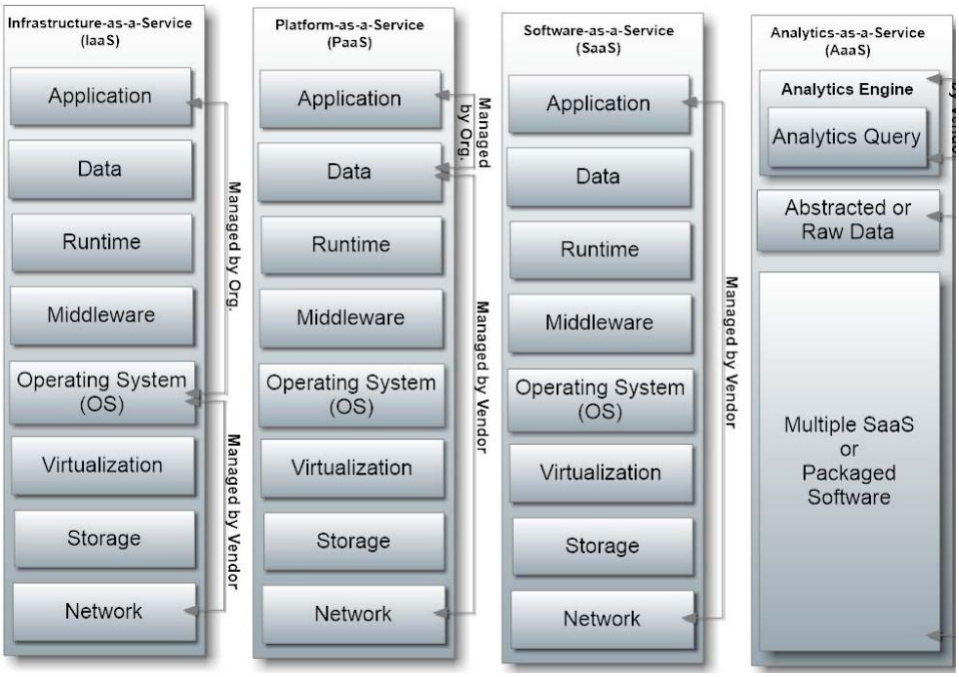

Figura 3: Modelos de servicio Cloud Computing revelantes

Dentro de Cloud Computing se encuentran las siguientes variantes:

- IaaS (Infrastructure as a Service)
  Se basa en el principio de que las inversiones en Hardware se hacen obsoletas y que las organizaciones no tienen que asumir dichos gastos por adelantado, en lugar de eso, las organizaciones rentan nubes de servidores que pueden prestar tiempo de ejecución, middleware, datos y aplicaciones dejando los gastos de mantenimiento al proveedor. Cuando las aplicaciones necesitan más recursos, la infraestructura que la soporta puede crecer tanto como se necesite [27].

- PaaS (Platform as a Service)
  Usualmente representado en los diagramas de Cloud Computing entre las capas SaaS e IaaS, es una amplia colección de aplicaciones de infraestructura (middleware) y servicios (incluyendo plataforma de aplicación, integración, gestión de procesos de negocio y servicios de bases e datos) [9].

- SaaS (Software as a Service)
  Software que es propiedad, distribuido y gestionado remotamente por uno o más proveedores. El proveedor entrega el software basado en un conjunto común de código y de definiciones de datos que es consumido en un modelo de uno-a-muchos por todos los clientes contratados en cualquier tiempo con una base de pago-por-uso o una suscripción basada en uso de métricas [10].

- AaaS (Analytics as a Service)
  El motor de análisis(compuesto por algoritmos y un procesador de consultas) es configurado por el proveedor, mientras que el resto de los datos transaccionales, software requerido, plataforma e infraestructura está en la organización. El proveedor de AaaS generalmente aloja el motor de análisis en IaaS y realiza los análisis de manera rápida, posiblemente co-relacionando con datos externos en un dominio público y entrega los resultados a la organización [27].

1.2.2. Analytics

Es la ciencia de examinar datos en bruto con el propósito de obtener conclusiones acerca de dicha información. Es usado en muchas industrias para permitirle a las compañías y organizaciones tomar mejores decisiones y en ciencias para verificar o desaprobar modelos existentes o teorías. Analytics

se distingue de la minería de datos por el alcance, propósito y foco del análisis. Los mineros de datos trabajan con grandes conjuntos de datos usando software sofisticado para identificar patrones no descubiertos y establecer relaciones ocultas, mientras que en Analytics el foco es en la inferencia, el proceso de llegar a una conclusión, basados solamente en lo que ya es conocido por el investigador [31].

Esta ciencia es generalmente dividida en Análisis Exploratorio de Datos (EDA), donde son descubiertas nuevas características en los datos y Análisis Confirmatorio de los Datos (CDA) donde las hipótesis existentes son probadas como verdaderas o falsas. Existe también el análisis Cualitativo de Datos (QDA) que es usado en las ciencias sociales para llegar a conclusiones basadas en datos no numéricos como palabras, fotografías o videos [31].

1.2.3. MapReduce

Es un estilo de computación que se puede usar para gestionar muchos cálculos de gran escala en una manera que es tolerante a fallos de hardware [28]. Se compone de dos tareas, la primera es la Map, que toma un conjunto de datos y lo convierte en otro donde los elementos individuales se dividen en tuplas y la tarea Reduce que toma la salida de la tarea Map como entrada y combina aquellos datos en un conjunto más pequeño de tuplas [33].

1.2.4. NoSQL

También llamado Not Only SQL, es un enfoque de gestión de datos y diseño de bases de datos que es muy útil para conjuntos de datos muy grandes y que se encuentran distribuidos. Busca resolver los problemas de escalabilidad y rendimiento que trae Big Data y para los cuales las bases de datos relacionales no fueron diseñadas para manejar. NoSQL es especialmente útil cuando una empresa necesita acceder y analizar montos masivos de datos no estructurados o que se encuentran almacenados remotamente en múltiples servidores virtuales en la nube. Una base de datos NoSQL puede organizar los datos en objetos, pares llave/valor o tuplas. Contrario a lo que dice su nombre, NoSQL no prohíbe el uso de el lenguaje estructurado de consultas (SQL). La base de datos NoSQL más popular es Apache Cassandra, seguido de otras como SimpleDB, Google BigTable, MongoDB0, HBase, MemcacheDB y Voldemort [32].

1.2.5. Apache Hive

Apache Hive provee un dialecto basado en SQL, llamado Hive Query Language(HiveQL) para la consulta de datos almacenados en un clúster Hadoop. Hive traduce las consultas a trabajos MapReduce, explotando la escalabilidad de Hadoop, a la vez que presenta una abstracción SQL fácil de aprender. Sin embargo Hive es mucho más apropiado para aplicaciones de bodegas de datos, donde se analizan datos relativamente estáticos, no se requieren tiempos de respuesta rápidos y los datos no están cambiando rápidamente [4].

1.2.6. Apache Pig

Apache Pig provee un motor para la ejecución de flujos de datos en paralelo sobre Hadoop y un lenguaje llamado Pig Latin para la expresión de dichos flujos de datos. Dicho lenguaje incluye operadores para muchas de las operaciones de datos tradicionales (join, sort, filter, etc.), como también ofrece la posibilidad de que los usuarios puedan desarrollar sus propias funciones para lectura, procesamiento y escritura de datos. Dado que pig corre sobre Hadoop, hace uso directo de HDFS (Hadoop Distributed File System)[4] Y MapReduce (El sistema de procesamiento de Hadoop) [12].

1.3. Distribuciones de Big Data

Teniendo en cuenta las tecnologías asociadas al paradigma de Big Data mencionadas en la sección 1.2, se muestran a continuación algunas de las implementaciones más importantes del mercado:

1.3.1. Hadoop

Es un *framework* que permite el procesamiento distribuido de conjuntos grandes de datos a través de clusters de computadores usando modelos de programación simples. Está diseñado para ir escalando de un solo servidor a miles de máquinas, cada una ofreciendo capacidad de cálculo y almacenamiento. En lugar de confiar en el Hardware para entregar alta disponibilidad, la librería por sí misma está diseñada para detectar y manejar fallos en la capa de aplicación, entregando así un servicio altamente disponible por encima de un clúster de computadores, cada uno de los cuales puede ser propenso a fallos [2].

[4]HDFS es un sistema de ficheros distribuido que almacena los archivos a través de todos los nodos en un cluster Hadoop

El *framework* incluye los siguientes módulos:

- Hadoop Common: Los utilitarios comunes que soportan los demás módulos de Hadoop.
- Hadoop Distributed File System (HDFS): Un sistema de archivos distribuido que provee acceso de alto rendimiento a datos de aplicación.
- Hadoop YARN: Un *framework* para la planificación de tareas y gestión de recursos en clúster
- Hadoop MapReduce: Un sistema basado en YARN para el procesamiento en paralelo de grandes conjuntos de datos.

1.3.2. Hortonworks

Hortonworks es un producto de código abierto que se enfoca en llevar al ambiente empresarial, de una manera mucho más transparente, la integración de las tecnologías de Big Data desarrolladas por Apache Software Foundation [14]. A continuación, en la figura4, se muestra la arquitectura de Hortonworks, en donde se puede ver la integración de las diversas tecnologías de Apache.

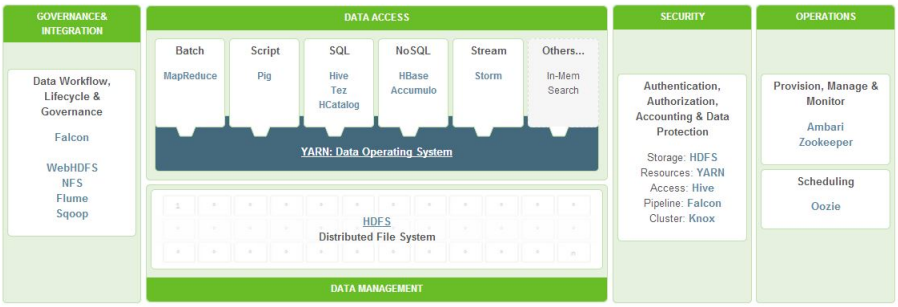

Figura 4: Arquitectura de Hortonworks

1.3.3. Pivotal Greenplum

Pivotal Greenplum [25] es un producto que propone la integración de tres elementos:

- Una base de datos de procesamiento paralelo para datos estructurados

- Una distribución Hadoop llamada Greenplum HD
- Chorus, una plataforma de productividad que incrementa la agilidad en análisis de equipos globales dedicados a la ciencia de datos. Provee una única interfaz para todos los datos de una organización con bases de datos virtuales, esto con el fin de facilitar la exploración, innovación y colaboración social para el análisis e identificación de oportunidades. Le permite a los analistas, personal de TI y ejecutivos participar en análisis Big Data [26].

1.3.4. IBM InfoSphere

IBM provee un portafolio de productos para Big Data como son [15]

- InfoSphere Streams: Permite el análisis continuo de grandes volúmenes de datos en streaming con tiempos de respuesta en sub-milisegundos.
- InfoSphere BigInsights: Una solución basada en Apache Hadoop lista para la empresa que permite la gestión y análisis de grandes volúmenes de datos estructurados y no estructurados.
- InfoSphere Data Explorer: Software de descubrimiento y navegación, el cual provee acceso en tiempo real y fusión de Big Data con datos ricos y variados de las aplicaciones empresariales para un mayor retorno de la inversión.

1.3.5. Microsoft HDInsight

Hace disponible Apache Hadoop como un servicio en la nube, haciendo el *framework* MapReduce disponible de una manera más simple, escalable y eficiente en términos de costo al soportarse en un ambiente Windows Azure. HDInsight también provee un enfoque eficiente en términos de costo para la gestión y almacenamiento de datos usado Windows Azure Blob [21].

1.3.6. Oracle Big Data

Ofrece una solución completa e integrada para abarcar el espectro completo de los requerimientos empresariales de Big Data. La estrategia en Oracle se centra en la idea de poder extender la arquitectura actual de información empresarial para incorporar Big Data, de tal manera que las nuevas tecnologías como Hadoop o la base de datos Oracle NoSQL, puedan trabajar en conjunto con una bodega de datos Oracle con el fin de entregar valor al

negocio [24]. En la figura 5 se presenta la manera en que se orquestan las soluciones de Big Data provistas por Oracle.

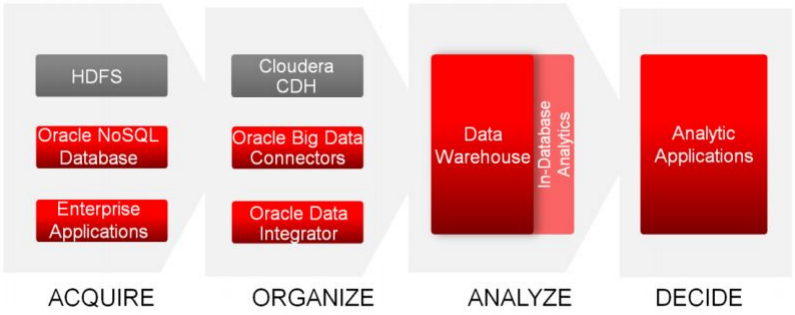

Figura 5: Soluciones de Big Data de Oracle

En el siguiente cuadro 3 se muestra un resumen comparativo con los diferentes productos relacionados a Big Data:

Tipo	**Producto**	**Base de datos**	**Implementación Hadoop**	**Componente NoSQL**
Open Source	Apache Hadoop	HBase	Hadoop	Hbase
Open Source	Hortonworks	HBase	Hadoop	Hbase
Comercial	Pivotal Greenplum	Greenplum	Greenplum HD	Hbase
Comercial	IBM Infosphere	DB2	InfoSphere BigInsights	HBase
Comercial	Microsoft BigData HD Insights	SQL Server	BigData Solution	SQL Server
Comercial	Oracle Big Data	Oracle	Cloudera	Oracle NoSQL

Cuadro 3: Productos de Big Data

1.4. Arquitectura de referencia de Big Data

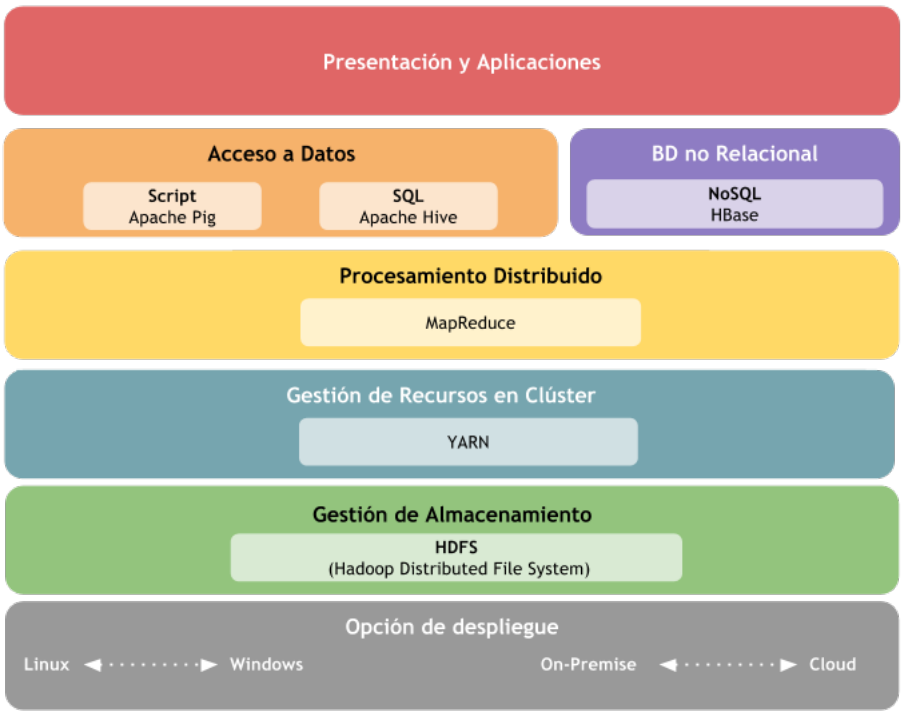

Figura 6: Arquitectura de referencia de Big Data

La arquitectura de referencia propone los siguientes componentes:

1. Opción de despliegue
   Son las diferentes opciones que se tienen para desplegar la arquitectura de Big Data, teniendo por ejemplo sistemas operativos que van desde distribuciones Linux hasta Windows, así como también ambientes de instalación que pueden ser On-Premise (en las instalaciones de la organización) o en la nube.

2. Gestión de almacenamiento
   Es la capa que almacena y procesa los datos de una manera distribuida, es representada por HDFS (Hadoop Distriuted File System). Se encarga de dividir el clúster en pequeñas piezas (llamados bloques) lo que permite a las funciones map y reduce ser ejecutadas en pequeños

subconjuntos de datos, lo que provee la escalabilidad necesaria para el procesamiento de Big Data [14].

3. Gestión de recursos en clúster
   Provee la gestión de recursos para permitir una amplia variedad de métodos de acceso a datos para operar con la información almacenada en HDFS con niveles de servicio y rendimiento predecibles. Esta capa está representada por Apache Hadoop YARN, el cual nace desde la versión 2.0 de Hadoop y su función es separar la gestión de recursos de los componentes de procesamiento [14].

4. Procesamiento distribuido
   Es el componente que mediante las fases de map y reduce permite el procesamiento en paralelo.

5. Acceso a datos
   Son los mecanismos mediante los cuales se puede acceder de manera simultánea a un conjunto de datos previamente distribuido. Se tienen entonces varias opciones, dentro de las que se encuentran Apache Pig y Apache Hive. Hive es mucho más adecuado para aplicaciones de bodegas de datos que son implementadas en bases de datos relacionales, reduciendo el impacto de migración y la curva de aprendizaje de las personas que ya manejan el lenguaje SQL. Sus desventajas son que las consultas tienen mayor latencia debido a la sobrecarga que genera el arranque de los trabajos MapReduce, no es transaccional y no ofrece tareas de actualización, inserción o eliminación a nivel de registros [4]. Por otro lado, Apache Pig ofrece un mayor nivel de abstracción de las tareas MapReduce (map, shuffle y reduce) en forma de un lenguaje de scripting (PigLatin) y a diferencia de HiveQL no es un lenguaje de consultas, sino que le ofrece la oportunidad al usuario de describir exactamente cómo procesar los datos de entrada. Pig Latin ofrece todas las operaciones de procesamiento estándar de MapReduce (join, filter, group by, order, union, etc.) y en general escribir programas en Pig Latin es mucho menos costoso de escribir y mantener que el código Java para MapReduce [12]. Pig además ofrece la posibilidad de crear funciones de procesamiento propias a través de las UDFs(User Defined Functions) que se pueden escribir en Java y en Python.

6. Base de datos no relacional
   Se puede también ver como una forma de acceso a datos pero con la particularidad de que usa HBase que es un sistema de gestión de

base de datos orientado a columnas que se ejecuta sobre HDFS. Hay que tener en cuenta que HBase no soporta SQL ya que no es relacional y las aplicaciones son escritas en Java de una manera parecida a MapReduce [14].

7. Presentación y aplicaciones
   Es la capa en la que se encuentran las aplicaciones existentes o nuevas que necesitan acceder a las bondades de la arquitectura Big Data propuesta por medio de APIs y servicios como REST.

2. Minería de texto

Luego de revisar las tecnologías y herramientas relacionadas a Big Data y teniendo en cuenta que uno de los grandes retos en el dominio del cuidado de la salud es el análisis del texto narrativo consignado en las historias clínicas electrónicas, se pasa a abordar las técnicas de minería de texto por ser las más indicadas para el análisis de dicho contenido [7].

La minería de texto puede ser ampliamente definida como un proceso de conocimiento intensivo en el que un usuario interactúa con un grupo de documentos a través del tiempo mediante el uso de un conjunto de herramientas de análisis. De una manera análoga a la minería de datos, la minería de texto pretende extraer información útil a partir de fuentes de datos a través de la identificación y exploración de patrones interesantes. En el caso de la minería de texto, sin embargo, las fuentes de datos son colecciones de documentos, y los patrones interesantes se encuentran no entre los registros de bases de datos sino en los datos de texto no estructurados en los documentos en dichas colecciones [7].

Como se muestra en la figura 7, un proceso simple de minería de texto arranca con un conjunto de documentos de texto, a los cuales se les aplican unas tareas de pre-procesamiento, obteniendo una colección de elementos que son la entrada a las operaciones núcleo de minería de texto cuyo resultado es presentado al usuario final.

Algunas de las tareas más críticas relacionadas a la minería de texto son el pre-procesamiento de colecciones de documentos (categorización de texto, extracción de información, extracción de términos), el almacenamiento de representaciones intermedias, técnicas para analizar dichas representaciones intermedias (como análisis de distribución, clustering, análisis de tendencias y reglas de asociación), así como la visualización de los resultados [7].

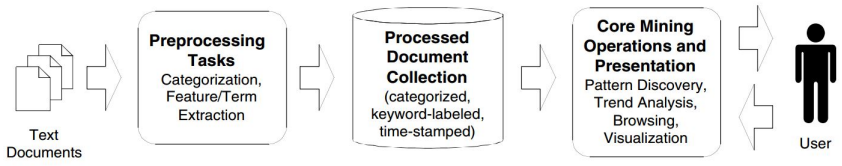

Figura 7: Arquitectura de alto nivel de un sistema de minería de texto [7]

Las operaciones de pre-procesamiento se centran en la identificación y extracción de características representativas para documentos en lenguaje natural y son responsables de la transformación de datos no estructurados almacenados en colecciones de documentos a un formato intermedio explícitamente estructurado [7].

Teniendo en cuenta que tanto las tareas de pre-procesamiento como las del núcleo son las dos áreas más críticas para cualquier sistema de minería [7] se encuentra allí una gran oportunidad de optimización de procesos, utilizando tecnologías de Big Data que permitan mejorar los tiempos de respuesta.

2.1. Técnicas de pre-procesamiento en minería de texto

A continuación se presentan algunas de las técnicas más importantes para el preprocesamiento en minería de texto:

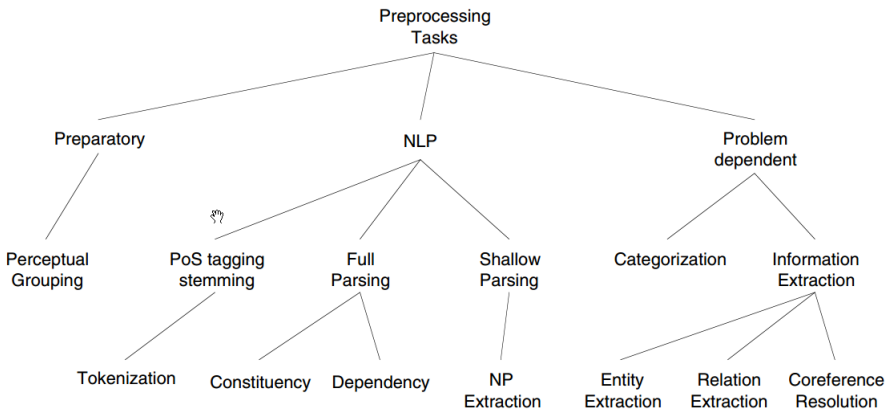

Figura 8: Taxonomía de tareas de preprocesamiento de texto [7]

2.1.1. Tokenization

Antes que cualquier otro procesamiento sofisticado, el flujo continuo de caracteres debe ser particionado en componentes significativos. Dicho proceso puede ocurrir a diferentes niveles. Los documentos pueden ser divididos por caractéres, secciones, párrafos, frases, palabras o incluso sílabas y fonemas.

El enfoque que se encuentra más frecuentemente en los sistemas de minería de texto incluye la división de texto en frases y palabras, lo cual es llamado tokenización. El principal desafío en la identificación de límites de la frase es distinguir entre un período que señala el final de una frase y un periodo que es parte de un token anterior como el Sr., Dr., etc. [7].

2.1.2. Part-of-speech Tagging

El etiquetado POS es la anotación de palabras con la etiqueta apropiada basado en el contexto en el cual ellas aparecen. Las etiquetas POS dividen las palabras en categorías basadas en el rol que ellas juegan en la frase en que aparecen [7]. Proveen información acerca del contenido semántico de una palabra. Los sustantivos denotan por lo general "las cosas tangibles e intangibles", mientras que las preposiciones expresan relaciones entre "cosas". La mayoría de los conjuntos de etiquetas POS hacen uso de las mismas categorías básicas. El conjunto más común de etiquetas contiene siete etiquetas diferentes (artículo, sustantivo, verbo, adjetivo, preposición, número y nombre propio). Algunos sistemas contienen un conjunto de etiquetas mucho más elaborado. Por ejemplo, el conjunto completo de etiquetas Brown Corpus tiene 87 etiquetas básicas. Por lo general, los etiquetadores POS en alguna etapa de su procesamiento realizan análisis morfológico de las palabras. Por lo tanto, una salida adicional de un etiquetador de este tipo es una secuencia de lemas de las palabras de entrada [7].

2.1.3. Syntactical Parsing

Los componentes de análisis sintáctico realizan una revisión completa de las frases de acuerdo a cierta teoría gramática. La división básica es entre las gramáticas de dependencia y de constituyentes [7].

La gramática de constituyentes describe la estructura sintáctica de las oraciones en términos de frases construidas de manera recursiva (secuencias de elementos agrupados sintácticamente). La mayoría de las gramáticas de constituyentes distinguen entre frases de sustantivos, de verbos, preposicionales, de adjetivos y cláusulas. Cada oración puede consistir en frases más

pequeñas o simplemente palabras de acuerdo a las reglas de la gramática. Adicionalmente, la estructura sintáctica de oraciones incluye los roles de las diferentes frases. Por lo tanto, una oración de sustantivo puede ser etiquetada como el sujeto de una oración, su objeto directo o el complemento [7].

La gramática de dependencias, por otro lado, no reconoce los constituyentes como unidades linguísticas separadas sino que se enfoca en las relaciones directas entre frases. Por ejemplo, un sujeto y sustantivos directos de una oración típica dependen del verbo principal, un adjetivo depende del sustantivo que modifica y así sucesivamente [7].

2.1.4. Shallow Parsing

El análisis eficiente y preciso de texto sin restricciones no se encuentra dentro del alcance de las técnicas actuales. Los algoritmos estándar son muy costosos para el uso en corpus[5] muy grandes y no son lo suficientemente robustos. el análisis sintáctico superficial (Shallow Parsing) compromete velocidad y robustez en el procesamiento al sacrificar la profundidad del análisis. En lugar de proveer un análisis completo de toda una frase, se producen sólo partes que son fáciles y sin ambigüedades. Típicamente, frases pequeñas de sustantivos y verbos son generadas, mientras que las cláusulas complejas no son formadas. Del mismo modo, se podrían formar la mayoría de las dependencias importantes, pero las poco claras y ambiguas se dejan sin resolver. Para los fines de la extracción de información, el análisis sintáctico superficial es por lo general suficiente y por lo tanto preferible al análisis completo debido a su superioridad en velocidad y robustez [7].

2.1.5. Categorización

Las tareas de categorización de texto (a veces llamado clasificación de texto) etiquetan cada documento con un número pequeño de conceptos o palabras clave. El conjunto de todos los posibles conceptos o palabras clave es normalmente preparado manualmente, cerrado y comparativamente pequeño. La relación de jerarquía entre las palabras clave también es preparada de forma manual [7].

[5] Colección de documentos

2.1.6. Extracción de informacion

La Extracción de información es quizás la técnica más prominente actualmente usada en las operaciones de preprocesamiento en minería de texto. Sin dicha técnica, los sistemas de minería de texto podrían tener unas capacidades de descubrimiento de conocimiento mucho más limitadas. Esta técnica debe ser distinguida de la recuperación de información o lo que se llama de manera informal "Búsqueda" ya que IR retorna documentos que coinciden con una consulta dada pero aún requiere que el usuario lea dichos documentos para localizar la información relevante. EI, por otro lado, tiene como objetivo la localización de información relevante y presentarla en un formato estructurado (típicamente en un formato tabular). Para los analistas y otros trabajadores del conocimiento, EI puede ahorrar tiempo valioso al acelerar dramáticamente el trabajo de descubrimiento [7].

Tanto las tareas de categorización de texto como de Extracción de Información le permite a los usuarios ir de una representación de los documentos leible por una máquina a una entendible por una máquina como se muestra en la figura 9.

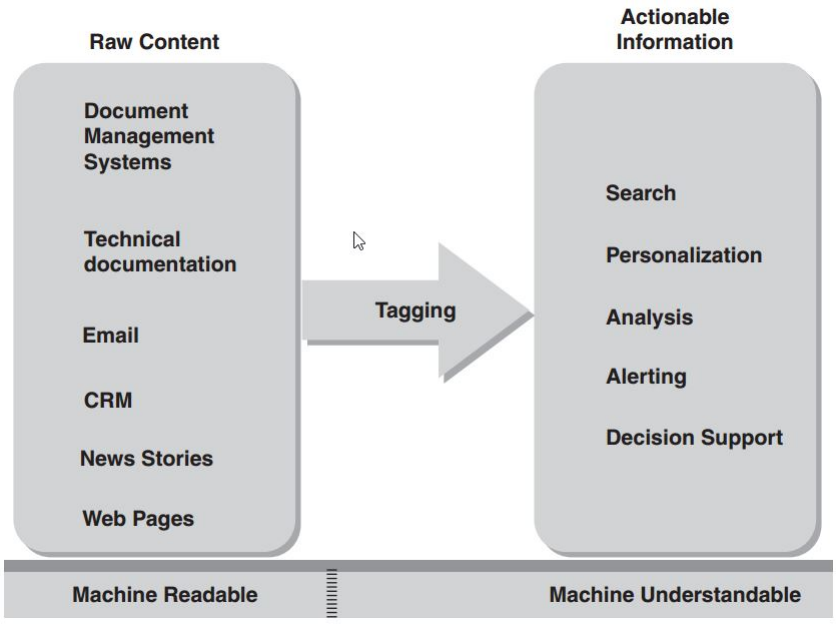

Figura 9: Cerrar la brecha entre los datos brutos e información procesable [7]

Luego de identificar las técnicas de pre-procesamiento en minería de texto, se pasa a la revisión de los casos de uso de Big Data en conjunto a Minería de texto en las bases de datos científicas.

2.2. Aplicaciones de Big Data en Minería de Texto

2.2.1. Big Data para análisis de datos no estructurados

Se propone la utilización de Big Data unido a minería de texto para trabajar con contenido narrativo en forma de tweets públicos, usando tecnologías como HBase, Java, REST y Hadoop [6].

Para ello se establecieron las siguientes fases:

1. El establecimiento de la conexión, seguida por la transmisión de los tweets públicos de Twitter utilizando Java (los datos se recuperan a partir de análisis de respuesta XML).

2. La construcción de un Hbase y almacenamiento de los datos en él, después del análisis de sentimiento (toda la comunicación se realiza a través de llamadas REST).

3. Construcción de un Front-end para que el cliente interactúe con la Hbase a través del *framework* Java para obtener los datos pertinentes requeridos

Dado que es un proyecto en progreso, solamente se ha completado la primera fase, en donde se toman los tweets y los colocan en la base de datos HBase, pero no se realiza minería de texto aún.

2.2.2. GATECloud.net: una plataforma de código abierto para el procesamiento de texto a gran escala en la nube

Se mencionan varias estrategias para mejorar los tiempos de respuesta en procesamiento de lenguaje natural, entre las cuales se tenían dos opciones: el uso de MapReduce y el manejo de tecnologías de Cloud Computing, siendo para ellos mejor la segunda, bajo un modelo PaaS(Platform as a Service), ya que no es necesario reescribir los algoritmos de procesamiento de texto que se van a ejecutar. El nombre de la solución es GATECloud.net, que es una plataforma Web en la que los científicos pueden realizar sus experimentos de minería de texto, cuyos servicios son cobrados mediante el modelo pago por uso y donde no se proponen librerías(APIs) para el empleo de la herramienta desde otros sistemas [30].

2.3. Herramientas de Minería de texto

A continuación se enuncia un conjunto de herramientas que son usadas para minería de texto:

2.3.1. Gate

Gate es un software de código abierto capaz de resolver problemas de procesamiento de texto, contando con una comunidad madura y extensa de desarrolladores, usuarios, educadores, estudiantes y científicos. Provee un proceso definido y repetible para la creación de flujos de trabajo para el procesamiento de texto, de tal forma que dichos flujos puedan ser mantenibles y robustos [11].

2.3.2. KNIME Text Processing

Permite leer, procesar, minar y visualizar datos textuales en una manera conveniente. Provee funcionalidades para [16]

- Procesamiento de Lenguaje Natural (NLP por sus siglas en inglés)
- Minería de texto
- Recuperación de información

2.3.3. Apache UIMA

UIMA(Unstructured Information Management Aplications) son sistemas de software que analizan grandes volúmenes de información no estructurada con el fin de descubrir conocimiento que es relevante a un usuario final. Adicional a ello, provee capacidades para envolver componentes como servicios de red y puede ser escalado a realmente grandes volúmenes por medio de replicación de *pipelines* de procesamiento sobre un clúster de nodos interconectados [1].

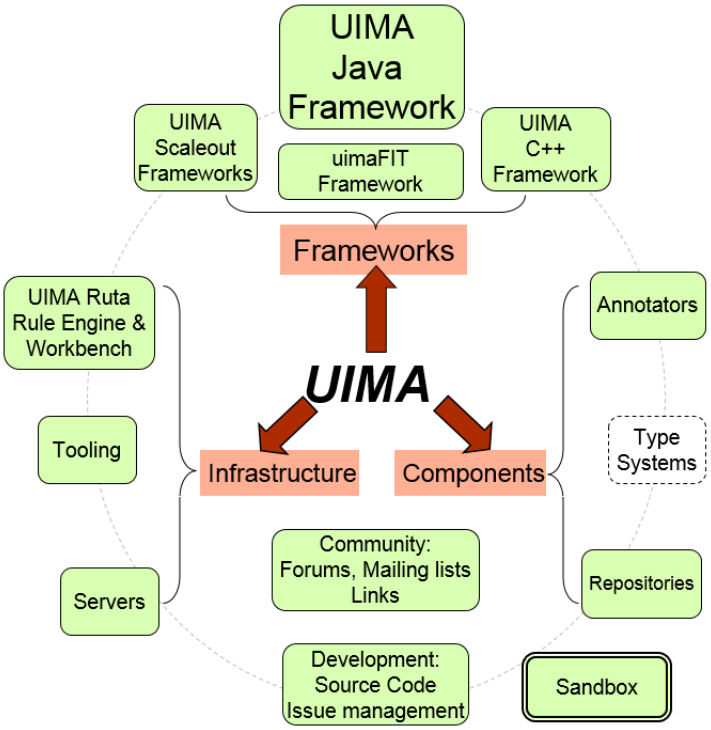

Figura 10: Arquitectura UIMA

2.3.4. SAS Enterprise Miner

El software SAS Enerprise Miner de texto (SAS Institute) es utilizado para el análisis de texto y su proceso de minería de texto incluye convertir texto no estructurado en datos estructurados, realizar clustering de las entradas de documentos/datos y ver los enlaces entre conceptos. Utiliza la técnica TF-IDF para identificar algunos de los conceptos o términos importantes [13].

Referencias

[1] Apache UIMA - apache UIMA. http://uima.apache.org/.

[2] Apache Software Foundation. Hadoop. http://hadoop.apache.org/.

[3] Benjamin H. Brinkmann, Mark R. Bower, Keith A. Stengel, Gregory A. Worrell, and Matt Stead. Large-scale electrophysiology: Acquisition, compression, encryption, and storage of big data. *Journal of Neuroscience Methods*, 180(1):185–192, May 2009.

[4] Edward Capriolo, Dean Wampler, and Jason Rutherglen. *Programming Hive*. O'Reilly Media, 1 edition edition, September 2012.

[5] Nitesh V. Chawla and Darcy A. Davis. Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, 28(S3):660–665, June 2013.

[6] T.K. Das and P Mohan Kumar. Big data analytics: A framework for unstructured data analysis. *School of Information Technology and Engineering, VIT University*.

[7] Ronen Feldman and James Sanger. *The Text Mining Handbook*. Cambridge University Press, 1 edition edition, March 2013.

[8] Bill Fox. Using big data for big impact. leveraging data and analytics provides the foundation for rethinking how to impact patient behavior. *Health management technology*, 32(11):16, November 2011. PMID: 22141243.

[9] Gartner. Platform as a service (PaaS). http://www.gartner.com/it-glossary/platform-as-a-service-paas.

[10] Gartner. Software as a service (SaaS). http://www.gartner.com/it-glossary/software-as-a-service-saas/.

[11] Gate. GATE. https://gate.ac.uk/.

[12] Alan Gates. *Programming Pig*. O'Reilly Media, 1 edition edition, September 2011.

[13] Claudia Gomez. Great: Modelo para la detección automática de relaciones semánticas entre resúmenes de texto provenientes de bases de datos del sector salud. http://pegasus.javeriana.edu.co/ PA123-03-SemantDatoSalud/anexos.html.

[14] Hortonworks. HDP 2.0 - the complete hadoop 2.0 distribution for the enterprise. http://hortonworks.com/products/hdp/.

[15] IBM. Big data portfolio of products. http://www-01.ibm.com/software/data/bigdata/platform/product.html.

[16] KNIMEtech. KNIME text processing. http://tech.knime.org/knime-text-processing.

[17] Harshana Liyanage, Siaw-Teng Liaw, and Simon de Lusignan. Accelerating the development of an information ecosystem in health care, by stimulating the growth of safe intermediate processing of health information (IPHI). *Informatics in primary care*, 20(2):81–86, 2012. PMID: 23710772.

[18] Viktor Mayer-Schonberger and Kenneth Cukier. *Big data: a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, Boston, 2013.

[19] Andrew McAfee and Erik Brynjolfsson. Big data: The management revolution. *Harvard business review*, 90(10):p60–68.

[20] William Q. Meeker and Yili Hong. Reliability meets big data: Opportunities and challenges. *Quality Engineering*, 26(1):102–116, January 2014.

[21] Microsoft. HDInsight. http://www.windowsazure.com/en-us/documentation/articles/hdinsight-get-started/.

[22] Keith D Moore, Katherine Eyestone, and Dean C Coddington. The big deal about big data. *Healthcare financial management: journal of the Healthcare Financial Management Association*, 67(8):60–66, 68, August 2013. PMID: 23957187.

[23] Travis B. Murdoch. The inevitable application of big data to health care. *JAMA*, 309(13):1351, April 2013.

[24] Oracle. Big data and oracle. http://www.oracle.com/us/technologies/big-data/index.html.

[25] Pivotal. Big data. http://www.gopivotal.com/big-data.

[26] Pivotal. Chorus. http://www.gopivotal.com/big-data/pivotal-chorus.

[27] Saptarshi Purkayastha and Jørn Braa. Big data analytics for developing countries – using the cloud for operational BI in health. *The Electronic Journal of Information Systems in Developing Countries*, 59(0), October 2013.

[28] Anand Rajaraman and Jeffrey D Ullman. *Mining of massive datasets*. Cambridge University Press, New York, N.Y.; Cambridge, 2012.

[29] Partho P. Sengupta. Intelligent platforms for disease assessment. *JACC: Cardiovascular Imaging*, 6(11):1206–1211, November 2013.

[30] V. Tablan, I. Roberts, H. Cunningham, and K. Bontcheva. GATECloud.net: a platform for large-scale, open-source text processing on the cloud. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1983):20120071–20120071, December 2012.

[31] WhatIs.com. What is data analytics (DA)? http://searchdatamanagement.techtarget.com/definition/data-analytics.

[32] WhatIs.com. What is NoSQL (not only SQL)? http://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL.

[33] Paul Zikopoulos and Chris Eaton. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media, 1 edition edition, October 2011.