Big Data Analytics: Oportunidades, Retos y Tendencias

Luis F. Tabares, Jhonatan F. Hernández

Especialización en Procesos para el Desarrollo de Software, Universidad de San Buenaventura
Cali, Colombia

***Abstract*— El constante uso de las tecnologías ha traído consigo un crecimiento explosivo en la cantidad de datos, los cuales son generados a grandes velocidades y en distintos formatos. A partir de este aumento de información, se da la necesidad de extraer de ella, patrones, tendencias y/o conocimiento, de forma rápida y eficiente, para lo cual, los métodos tradicionales han tenido que evolucionar en busca de rendimiento y escalabilidad. El gran contenido de valor que genera este tipo de información está permitiendo a las organizaciones una mejora en la toma de sus decisiones, lo que conlleva a la obtención de ventajas competitivas en los diferentes campos de acción. En este artículo se describe el estado del arte, las oportunidades, retos y tendencias que existen sobre "Big Data Analytics", con un enfoque hacia el Software. Este artículo permitirá guiar futuras investigaciones y aplicaciones en esta área, de tal forma que pueda ser utilizado como referencia de línea base.**

Palabras Clave: *Big Data, Analytics, NoSQL, Cloud Computing, Data Warehouse, KDD*

I. INTRODUCCIÓN

El constante avance de las tecnologías ha permitido un crecimiento "explosivo" en la cantidad de datos generados desde diferentes fuentes, tales como, redes sociales, dispositivos móviles, sensores, máquinas de rayos x, telescopios, sondas espaciales, log de aplicativos, sistemas de predicción del clima, sistemas de geo-posicionamiento y, en términos generales, todo lo que se puede clasificar dentro de las definiciones del Internet de las Cosas [1]. De la mano de este considerable aumento de información, ha surgido la necesidad de extraer de ella, de manera eficiente, patrones, tendencias y/o conocimiento que permitan apoyar la toma de decisiones, para lo cual, los métodos tradicionales de procesamiento de datos han tenido que evolucionar rápidamente, buscando escalabilidad y rendimiento principalmente, con el fin de suministrar respuestas en tiempo real, al menor costo posible. A este fenómeno se le ha llamado Big Data y, según [2], [3], hace referencia principalmente a tres términos conocidos como las 3 Vs: Volumen, Velocidad y Variedad [2]. Pero Big Data no solo hace referencia a los problemas relacionados con los datos (enmarcados dentro de las 3 Vs), sino que también incluye un amplio espectro de técnicas, tecnologías, métodos y paradigmas no convencionales que apoyan la solución de problemas relacionados con datos de una forma diferente y, generalmente, más adecuada que los métodos tradicionales. Dado lo anterior, Big Data permitió entonces nuevas y mejores formas de procesar la información, con ventajas sobre los enfoques tradicionales, los cuales no responden de forma adecuada sobre las necesidades actuales de las compañías, en términos de velocidad, costos de implementación, escalabilidad, flexibilidad y elasticidad sobre entornos más complejos. Estos enfoques se encuentran orientados principalmente a la computación distribuida y el procesamiento paralelo masivo, que han convergido de cierta forma con tecnologías como la computación en la nube ("Cloud Computing") y las nuevas formas de almacenar los datos, mediante modelos no relacionales, sobre todo cuando se deben tener en cuenta los costos de dicho almacenamiento para su posterior procesamiento. Alrededor de estos paradigmas existen arquitecturas de referencia, patrones de diseño y tendencias en Software y Hardware encaminados a facilitar el uso de Big Data con el objetivo de generar ventajas competitivas y comprender el mundo de una forma más eficiente y eficaz. Grandes compañías de tecnologías de información como Google, Yahoo, Facebook y Amazon han investigado y se han apropiado de proyectos de gran relevancia y escala que han permitido, en principio, resolver los problemas inherentes a la gestión de Big Data. De estos proyectos surgieron modelos de almacenamiento distribuido de datos (como BigTable de Google [4], Dynamo de Amazon [5] y todas sus derivadas) y arquitecturas y algoritmos de procesamiento paralelo masivo (como MapReduce [6], Google File System [7], Apache Hadoop y Hadoop File System [8]), como parte de los nuevos métodos para trabajar con Big Data. Lo que resulta interesante es que estas tecnologías, modelos, técnicas de procesamiento de datos y servicios en la nube han sido utilizadas y apropiadas en otro tipo de compañías (por ejemplo, de comercio electrónico, sector gobierno/público y salud) que generan Big Data. Del mismo modo, la ciencia ha logrado sacar provecho de las mismas para realizar investigaciones y experimientos en diversas áreas de conocimiento (por ejemplo, física, bioinformática, astronomía y genética).

Pero los datos almacenados y gestionados no representan por sí solos la ventaja que requieren los diferentes sectores para ser más competitivos y productivos. Si bien, manejar Big Data adecuadamente en una organización (privada, gubernamental, o de cualquier sector) pudiera representar una ganancia, lo verdaderamente importante para esta organización es el **Valor** que se puede generar partir de estos datos, siendo esto aún más importante, si se parte de la premisa que indica que, en la mayoría de los casos, los datos no siempre generan este valor esperado, por lo que las organizaciones deben estar en la capacidad de descubrir esto en tiempo real sin descuidar aspectos de seguridad, integración, funcionalidad y otros atributos de calidad que sean contemplados en cada dominio en particular. Como ejemplos de los beneficios que se obtienen al

saber extraer dicho valor, se puede evidenciar, en el campo de la economía, que se ha logrado aumentar la productividad de las empresas mediante el entendimiento de sus nichos de mercado, a partir de Big Data; en el sector Gobierno, se ha logrado descubrir patrones demográficos a partir de redes sociales, diarios electrónicos y diferentes campañas que censan datos para la toma de decisiones de diferente índole; mientras que en el campo de la investigación científica se ha logrado analizar datos generados en diferentes ciencias y áreas de investigación (astronomía, meteorología, computación social y bioinformática) para obtener patrones y tendencias que han permitido entender procesos físicos, naturales, químicos, genéticos, entre otros. El presente artículo se enfoca principalmente en proporcionar un estado del arte de lo que se conoce como la cuarta V (Valor) o "Big Data Analytics", partiendo de la premisa de que el principal reto de esta área de estudio se encuentra en transformar la Big Data en conocimiento y llevar estas aplicaciones a las organizaciones, las cuales agregan retos adicionales como lo son: El costo computacional, la seguridad informática, la integración con otros sistemas, la volatilidad de los requisitos y demás aspectos propios de cada negocio o área de dominio. El artículo se encuentra organizado de la siguiente forma: En la sección II, se describe lo qué es Big Data tomando como referencia el fenómeno de las 3 Vs y cómo este representa una oportunidad para las organizaciones. En la sección III, se describe el enfoque o la idea que propone el presente artículo de estudiar Big Data desde el punto de vista de cómo puede esta ser analizada, para lo que se propone la adopción de una cuarta V por parte de los autores. En la sección IV, los autores suministran algunas ideas generales sobre las áreas de aplicación donde el análisis de Big Data tiene cabida. En las secciones V, VI y VII se describen los detalles de este enfoque de estudio, proporcionando: *(1)* Una definición de cuatro conceptos clave que giran en torno a Big Data Analytics (sección V); *(2)* La arquitectura general para analizar Big Data, adoptada por la mayoría de las organizaciones estudiadas (sección VI); y *(3)* Las oportunidades, retos y tendencias que se encuentran en curso, en relación con Big Data Analytics (sección VII). Finalmente, a partir de este estado del arte, se llegará a lo que podría constituirse como el trabajo futuro (sección VIII), donde se dará paso a una discusión final sobre los temas relevantes tratados y, de esta forma, buscar articular el resultado de éste con otros trabajos relacionados (seguimientos tecnológicos y bibliográficos, nuevas investigaciones o proyectos aplicados). En la última sección de este artículo (sección IX), se proporciona una conclusión final a partir de los resultados obtenidos en la investigación y los diferentes puntos de vista de los autores y organizaciones consultados. El principal objetivo del artículo es elaborar una línea base para apalancar, teóricamente, futuros trabajos sobre este campo, desmitificando algunos conceptos errados que generan sesgo sobre este tema.

II. BIG DATA: LAS 3 VS

El fenómeno que se presenta con Big Data se encuentra motivado principalmente por el advenimiento y/o crecimiento de olas tecnológicas que marcan una "revolución", como es el caso del Internet de las cosas (o IoT, por sus siglas en Inglés) [1]. Estos fenómenos hacen que se generen, a través de diversos medios, contenidos dispuestos en diferentes formatos (principalmente, texto, audio y video), registrados a grandes velocidades y accedidos frecuentemente en busca de información útil. Esto ha producido una "explosión de datos" conocida generalmente como "Big Data" [2]. Para ejemplificar lo anterior, se tiene que, según [9], la cantidad de datos digitales a nivel mundial creció de 150 exabytes (billones de gigabytes), en 2005, a 1.200 exabytes, en 2010. En 2007 se preveía que dicha cantidad aumentara en un 40% anual en los años siguientes (unas 40 veces el crecimiento de la población del mundo), por lo que se esperaba que la cantidad de datos digitales aumentara 44 veces entre 2007 y 2020 (duplicándose cada 20 meses). En la Figura 1, se puede ver cómo entre los años 2005 y 2011 la información ha crecido de forma exponencial con respecto al almacenamiento disponible. A lo anterior también se le ha conocido como la "revolución de los datos" y llega acompañada de nuevos tipos de problemas, retos y, sobretodo, oportunidades que se pueden abordar mediante el uso de nuevas herramientas dadas por técnicas y tecnologías alternativas a las tradicionales.

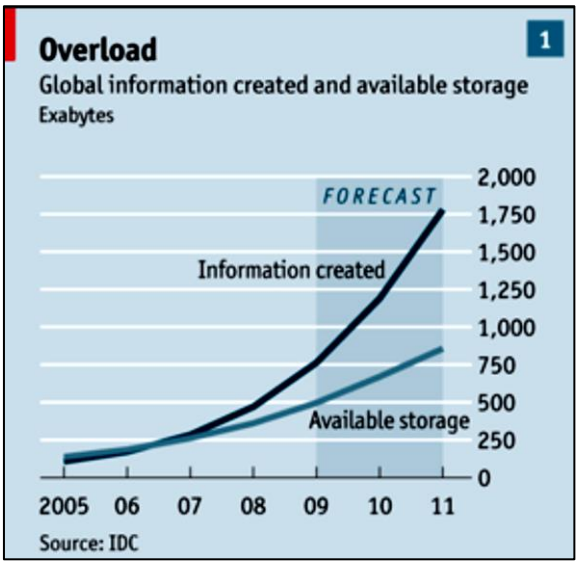

Figura 1: Los primeros años de la Revolución de los Datos. Tomado de [9]

Big Data es una nueva tendencia tecnológica que se ha sido estudiada por Gartner en [10], y es el foco de muchas investigaciones actuales y futuras. El concepto de Big Data es caracterizado en [2] por tres (3) términos o variables conocidas como las 3Vs: Volumen, Velocidad y Variedad (Figura 2), donde *Volumen* hace referencia a la cantidad de datos generados continuamente en un espacio de tiempo determinado; *Velocidad*, a la rapidez de entrada y salida con la cual éstos fluyen a través de distintos canales; y *Variedad*, a los diferentes formatos y fuentes en que éstos se encuentran. Las 3 Vs corresponden a la definición base de lo que representa Big Data para la mayoría de autores expertos en el tema. No obstante, según [3], Big Data no necesariamente involucra estas 3 Vs, teniendo así, que pueden identificarse entornos de Big Data, con la presencia de una o más Vs. También existen autores, sobretodo en la industria (como por ejemplo Teradata [11]), que defienden posiciones que definen a Big Data en

ausencia de las 3 Vs ocasionalmente. Por otro lado, según estos mismos autores, Big Data no solo hace referencia a los datos si no a las técnicas, métodos y tecnologías utilizadas para solucionar un problema, lo cual aleja un poco esta definición de las 3 Vs. En este orden de ideas, y para evitar generar una discusión en este artículo sobre qué es y qué no es Big Data, se puede decir, en términos generales, que se está abordando un problema relacionado con Big Data si éste requiere de técnicas o tecnologías diferentes a las tradicionales para resolverlo de una forma más adecuada, ya sea porque sus datos son complejos (3 Vs) o, simplemente, las técnicas y métodos tradicionales no son suficientes para dar una solución efectiva.

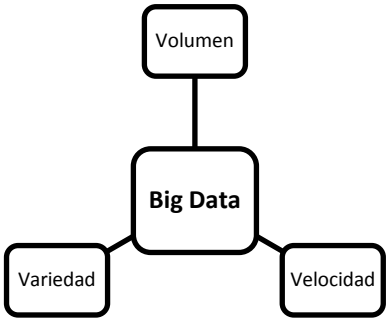

Figura 2. Big Data y las 3Vs (Adaptada de [2]**)**

Big Data tiene aplicación en diferentes dominios, tales como los descritos en [1]–[3], [9], [12]: Transporte y Logística, Salud, Ambientes Inteligentes, Computación Social y Personal, Ambientes Futurísticos, Gobierno, Comercio, Medio Ambiente e Investigación Científica. Esto hace que este nuevo paradigma constituya una importante oportunidad para entender el mundo desde los diferentes dominios o áreas de aplicación en donde se generan datos que pueden ser transformados en información útil para la toma de decisiones. Sin embargo, Big Data no solo trae consigo oportunidades y nuevas formas de solucionar los problemas computacionales. La "explosión de datos" descrita anteriormente también conlleva a nuevos desafíos (como la seguridad informática y la minimización de costos de almacenamiento y procesamiento) que obligan a dar una mirada más extensa y global a las alternativas que se proponen y, sobre todo a las necesidades puntuales de cada organización o unidad productiva.

III. BIG DATA ANALYTICS: LAS 4 VS

La mayoría de los autores se basan en las 3 Vs cuando se refieren a Big Data. Sin embargo, dado que este tema ha dado mucho de qué hablar y ha generado bastante publicidad e interés por parte de diferentes sectores, los enfoques han ido evolucionando al punto que los autores apropian más Vs en sus estudios o implementaciones sobre Big Data. Por ejemplo según [13], existen tres Vs adicionales: Veracidad, Validez y Volatilidad, donde *Veracidad*, hace referencia a los datos libres de ruido, a través de los cuales se puede hacer minería y análisis, es uno de los retos más grandes en Big Data; *Validez*, a la correctitud y exactitud de los datos para el uso que se quiere dar a los mismos; y *Volatilidad*, al tiempo que los datos serán almacenados y por cuánto tiempo serán válidos. Estas tres nuevas Vs dan paso al enfoque del presente artículo, el cual encierra estas nuevas variables en una conocida como la cuarta V (Figura 3), la cual se refiere al **Valor** que puede obtener a partir de los datos mediante procesos analíticos. Esta V puede considerarse como una de las más importantes y se presenta como la oportunidad de extraer información valiosa mediante el procesamiento de datos complejos, usando los Gigabytes, Terabytes o incluso Petabytes que las organizaciones generan, en forma de datos estructurados y no-estructurados, con la finalidad de convertirlos en decisiones y/o conocimiento. Una organización que no es capaz de entender sus datos y llevar a cabo acciones ejecutivas con éstos, no realmente productiva y tampoco podrá generar ventajas competitivas. En este tipo de empresas sucederá que los datos estarán allí simplemente almacenados sin generar mayores beneficios -por el contrario, aumentando los presupuestos de TI de manera inoficiosa-.

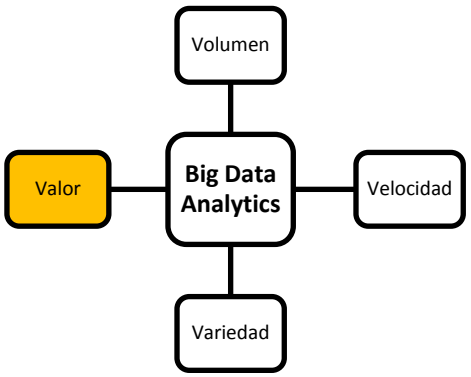

Figura 3: Big Data Analytics y las 4 Vs (Adaptada de [2]**)**

Durante el estudio de esta cuarta V se deben tener en cuenta la Veracidad, Validez y Volatilidad descritas anteriormente como variables importantes intrínsecas en el proceso de generar Valor, bien sea como parte de las oportunidades o los desafíos que esto representa. La relación entre estas características, representada en la Figura 4, puede explicarse de la siguiente forma: La naturaleza y procedencia de Big Data hace que los datos contengan ruido, el cual no garantiza la obtención de valor en dichos datos. La eliminación de este ruido proporciona *Veracidad* y, por ende se facilita la *Validez* de los mismos, de acuerdo con el propósito específico para el que son utilizados. Al mismo tiempo, cuando se genera valor es necesario conocer la *Volatilidad* de los datos para efectos de conocer los diferentes métodos de análisis a aplicar.

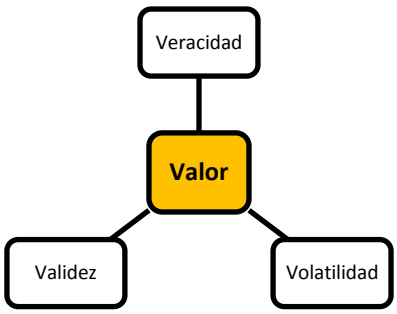

Figura 4: Valor y su relación con Veracidad, Validez y Volatilidad

En la actualidad, las organizaciones se encuentran más preocupadas por la información relevante que puede obtenerse a partir de los datos y es allí donde esta cuarta V toma mayor importancia. Ejemplos de lo anterior son:

- Una cadena de almacenes Retail quisiera conocer al instante lo que están comprando sus clientes, dado un evento o acontecimiento (como un fenómeno climático o de carácter social), a partir de los datos generados por su sistema de Puntos de Venta POS (Point Of Sales en inglés).
- Un centro de sismología podría predecir un terremoto horas antes de que éste ocurra a partir de datos -con ciertos patrones- provenientes de sensores.
- Una entidad de ayudas humanitarias debería poder mejorar la logística en una emergencia a partir de datos publicados en las redes sociales.
- Los diferentes Gobiernos necesitan comprender variables importantes como el clima y la demografía para incluir medidas y planes en sus políticas de gobierno.

Estas necesidades no podrían ser resueltas con el manejo de Big Data por sí solo. Para esto se requiere que los datos sean procesados y analizados en busca de conocimiento, tal como lo define el proceso del KDD [14], haciendo que sea necesario abordar arquitecturas y técnicas como el Data Warehousing. De lo anterior, y según diferentes autores, se puede definir entonces Big Data Analytics o análisis de Big Data, a partir de la cuarta V, como un amplio conjunto de métodos computacionales, estadísticos y de visualización utilizados para lograr una mejor comprensión de los datos en ambientes o entornos de Big Data, caracterizados por la complejidad, bien sea, de los datos o de los escenarios que se plantean a través de los diferentes dominios de aplicación.

IV. Aplicaciones de Big Data Analytics

Las áreas de aplicación de Big Data Analytics han sido las mismas durante varios años, con la diferencia de que hoy en día éstas pueden verse más beneficiadas por las ventajas que ofrece la revolución de los datos descrita en la sección II y el advenimiento de los nuevos métodos y técnicas de Big Data Analytics. Según [3], la productividad y competitividad de las empresas y la administración pública se pueden incrementar gracias a la Big Data. Según estos mismos autores -y los relacionados en [1], [2], [9], [15]–[19]-, en otros campos como las disciplinas científicas, la computación social y personal, el comercio y los negocios, el gobierno y la administración pública, la salud y el cuidado humano, los servicios públicos y la manufactura, existe una gran incidencia por parte del análisis de Big Data como acelerador de su desarrollo.

En términos generales, existe una gran variedad de áreas de aplicación que pueden beneficiarse del análisis de Big Data. En este artículo, se describen algunas de las más relevantes, que corresponden a las más nombradas por diferentes autores y representantes de la industria estudiosos del tema. En las siguientes sub secciones se abordan brevemente con el fin suministrar una idea general de estos beneficios. Al final de la sección, en la Figura 5, se muestra un resumen de éstas aplicaciones.

A. Disciplinas científicas

Ciencias y áreas de investigación como la astronomía, meteorología, biología, geología, oceanografía y sociología, están incrementando el uso de sensores para registrar altos volúmenes de datos heterogéneos. Muchos de estos datos también son simulados por científicos para fines investigativos, como es el caso del genoma humano, los aceleradores de partículas y el estudio del clima. El análisis de estos datos ha arrojado resultados invaluables que han permitido el entendimiento de múltiples procesos en distintos contextos, lo cual no era posible, con tal facilidad, hace algunos años.

B. Computación social y personal

Las redes sociales, blogs, publicaciones, comentarios, foros, búsquedas en Internet, el tráfico generado en los diferentes sitios web y demás acciones que se realizan de manera personal y/o social, están generando una extensa variedad de aplicaciones que van desde el anuncio inmediato de desastres (que permite la optimización de los niveles de atención suministrados por diferentes organizaciones), hasta la predicción de crímenes, fluctuaciones del mercado y análisis de sentimientos de las personas (para efectos de prevención y estrategias de marketing). La computación social es una de las principales motivaciones del surgimiento de análisis de Big Data y sus conceptos alrededor, atribuyéndosele gran parte de la "revolución" o "explosión" de los datos.

C. Sector comercio / negocios

La aplicación más visible del análisis de Big Data quizá se encuentre en este sector, en el cual se destacan, la industria Retail y el sector financiero, principalmente. En Retail se puede observar el beneficio que se obtiene de Big Data Analytics en lo que se conoce como la fidelización de clientes y el análisis de mercadeo, los cuales permiten crear estrategias de venta efectivas, basadas en las relaciones existentes en los diferentes objetos de negocio. El sector financiero toma provecho de Big Data Analytics en importantes implementaciones como la detección de fraudes y análisis de perfiles de clientes para su clasificación y posterior lanzamiento de estrategias de marketing y/o fidelización.

D. Gobierno y sector público

El sector público también involucra problemas de Big Data en la medida en que las poblaciones de las regiones (países, ciudades, comunidades) pueden ser grandes y con necesidades diversas o sectorizadas, en donde cada individuo genera millones de datos a través de diferentes canales, generalmente públicos (como los servicios públicos y las redes sociales, por ejemplo). Con Big Data Analytics es posible obtener resultados en tiempo real que permitan establecer políticas públicas que respondan a las diversas necesidades de manera efectiva y eficaz.

E. Sector salud

Es uno de los sectores más importantes para el desarrollo de las regiones, pero también es uno de los más complejos en términos de datos, ya que contempla información clínica, la cual, en un futuro podrá contener más datos de los que hoy se contemplan (por ejemplo, información genética); también contempla datos farmacéuticos, datos de prácticas, preferencias y hasta registros financieros de los pacientes. La integración de

todos estos datos será un factor clave para tener un sistema de salud mejor y más asequible (sin entrar en discusiones políticas). Con Big Data Analytics podrá ser posible tener una vista 360 grados no solo de pacientes sino también de organizaciones de salud, así como también se podrá optimizar el funcionamiento de las entidades que componen este sistema (por ejemplo, hospitales e intermediarios).

F. Servicios públicos y Telecomunicaciones

En la prestación de servicios públicos como el agua, gas, electricidad y telecomunicaciones, puede usarse Big Data Analytics como la base para la detección de filtraciones, fraudes y pérdidas no-técnicas, así como para realizar "mediciones inteligentes", partiendo de que los datos que son generados por los dispositivos y sistemas involucrados en la prestación de estos servicios son caracterizados por encontrarse en ambientes de Big Data.

G. Sector manufacturero

En este sector se pueden utilizar los datos procedentes de las máquinas de manufactura (generados por sensores, por ejemplo) para que, combinados con el efectivo análisis de la demanda, permitan una producción óptima, minimizando el desperdicio y el re-trabajo. Esto es posible a través de métodos relacionados con Big Data Analytics, que proporciona técnicas de avanzada para apoyar la toma de decisiones sobre estos datos caracterizados por ser complejos.

Disciplinas científicas	Entender procesos naturales
	Estudios sobre genética
	Simulaciones y experimentos
Computación social y personal	Optimizar atención a desastres
	Predicciones de diversa índole
	Análisis de sentimientos
	Análisis de diferentes fenómenos sociales y personales
Sector comercio / negocios	Marketing
	Gestión efectiva de clientes y sus relaciones
	Detección de fraudes
	Análisis financieros
Gobierno y sector público	Sectorización efectiva
	Medición de la adherencia de los ciudadanos a las leyes
	Control poblacional y migratorio
	Análisis de fenómenos políticos, sociales, culturales
Sector salud	Optimización de los servicios de salud
	Vista 360 grados de pacientes
	Detección de enfermedades y prevención
	Prevención y control de epidemias y pandemias
Servicios públicos	Optimizar mediciones y cobros
	Detectar fraudes y filtraciones
	Segmentación de clientes para personalizar servicios
	Monitoreo más proactivo
Sector manufactura	Optimizar el funcionamiento de las máquinas y uso de recursos
	Mayor proactividad con respecto a la demanda
	Mejorar ciclos de vida de producción
	Minimizar/evitar el desperdicio y el re-trabajo

Figura 5: Resumen Aplicaciones de Big Data Analytics

V. CONCEPTOS ALREDEDOR DE BIG DATA ANALYTICS

Alrededor de Big Data Analytics existen tres conceptos o variables principales, las cuales se encuentran estrechamente relacionadas entre sí: Data Warehouse, NoSQL y Cloud Computing (Figura 6). Algunas de estas variables pueden afectar a las otras en diferentes contextos y proporciones. Por ejemplo, en la medida que se generan más servicios que permitan la computación en la nube (Cloud Computing), se puede dar solución más fácilmente a problemas Big Data con costos más competitivos. A su vez, en la medida que estos problemas surgen, aparecen soluciones tecnológicas como NoSQL (y sus diferentes implementaciones) que resuelven el almacenamiento y gestión de los datos. Si a lo anterior se le da una connotación analítica (que corresponde al enfoque de este artículo), entonces surgirán también más soluciones de Inteligencia de Negocios en forma de aplicaciones, arquitecturas, técnicas, tecnologías, herramientas o nuevos paradigmas que apuntan a abordar el análisis de Big Data.

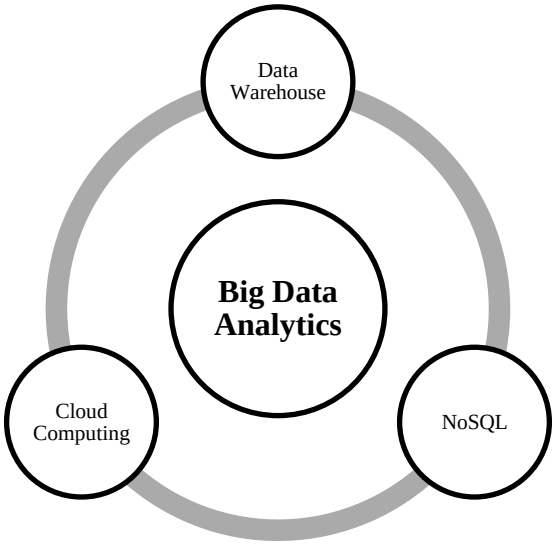

Figura 6. Tres conceptos alrededor de Big Data Analytics

Antes de discutir sobre Big Data Analytics, su estado del arte y los retos y oportunidades que brinda, es necesario comprender estos tres conceptos y el papel que éstos llevan a cabo en este enfoque. A continuación, se proporciona un marco sobre el cual se describen algunas definiciones que han adoptado diferentes autores y que han sido apropiadas en este artículo como fundamento teórico de aspectos clave que se nombran en el desarrollo del mismo.

A. Data Warehousing

Este artículo se refiere al Data Warehousing como el conjunto de arquitecturas que facilitan o apoyan procesos de KDD (Descubrimiento de conocimiento o Información valiosa a partir de los datos) o BI (Inteligencia de Negocios). Los conceptos sobre "Data Warehousing" comenzaron a surgir entre los años 70 y 80, con la necesidad de almacenar y analizar los datos generados a través del OLTP [2], después del bastante conocido artículo de Codd sobre su propuesta de un modelo relacional [20], el Sistema R [21] (que fue la primera Base de Datos relacional experimental) y la proliferación de

sistemas OLTP propuesta por el nuevo concepto relacional de aquel tiempo, el cual impuso –inclusive- cambios arquitectónicos y la posibilidad de desarrollar aplicativos más potentes para la época. Existen dos definiciones importantes sobre el proceso del KDD proporcionados en [14]: (1) "*Extracción no-trivial de información implícita, previamente desconocida y potencialmente útil, a partir de los datos*". (2) "*Exploración y análisis, automático o semi-automático, de grandes cantidades de datos para descubrir patrones significativos*". A partir de esta definición, se habla entonces de que la arquitectura principal que soporta este proceso es el Data Warehouse (Bodega o Almacén de datos), que es definido en [22] como "*Colección de datos en la cual, dos o más fuentes de datos dispares pueden ser reunidas mediante una estrategia de gestión integrada y variable en el tiempo*", en [23] como "*Colección de datos, orientada a temas, no volátil, integrada y variable en el tiempo, para la toma de decisiones*" y en [24] como "*Nada más que la unión de todos los data marts que lo constituyen*", siendo un Data Mart un proyecto de Data Warehouse enfocado a un proceso de negocio, el cual es más fácil de construir comparado con un Data Warehouse completo. En los procesos de Data Warehousing normal, una de las técnicas de análisis de información más utilizadas es el procesamiento por lotes (Batch), donde los datos generados en ventanas de tiempo son divididos en segmentos de tamaño fijo y procesados a través de diferentes capas, cuyo resultado es almacenado en bodegas de datos para su uso futuro en análisis y/o visualización. Con el tiempo, éste ha evolucionado en lo que conocemos hoy como su segunda generación, descrita en [23] y [25], donde se hace énfasis a la escalabilidad y flexibilidad, aún en cumplimiento de las reglas de Codd.

B. NoSQL

Las nuevas características que impone Big Data sobre los datos, hace que éstos se tornen imperfectos, complejos y sin estructura o sin esquema (Ejemplo: Archivos de texto, audio, video, colectores de datos, datos de sensores, huellas digitales, tweets, comentarios). Esto hace que deban aprovecharse herramientas computacionales avanzadas desarrolladas inclusive para otros campos (Machine Learning, algoritmos dividir y vencer, teoría de grafos y la utilización de mapas de datos). La mayoría de estas herramientas y tecnologías que tienen qué ver específicamente con el almacenamiento de los datos no se encuentran en cumplimiento de las reglas de Codd y se les conoce como NoSQL o no relaciones. Además de proponer diferentes modelos de datos, estas herramientas incluyen arquitecturas de computación distribuida y lenguajes de alto nivel para acceder a los datos, hecho que las ubica como las tecnologías y herramientas más utilizadas para Big Data.

Según [26]–[30], NoSQL se define como "*Término utilizado para describir una extensa variedad de tecnologías que proporcionan un enfoque alternativo para el almacenamiento de datos, comparado con los sistemas manejadores de bases de datos relacionales tradicionales*". Estas tecnologías se caracterizan por mantener una organización muy simple o sin esquema, con una semántica llave-valor y un esquema de almacenamiento altamente escalable. En general, estas bases de datos no ofrecen semántica SQL ni cumplen con las propiedades ACID (*Atomicity, Consistency, Isolation, Durability*), las cuales caracterizan al modelo relacional propuesto por Codd. En lugar de esto, las bases de datos NoSQL cumplen con las propiedades BASE (*Basically available, Soft State, Eventually consistent*) [26]. Uno de factores clave de las tecnologías NoSQL, es que los modelos de datos suministrados por éstas estructuran la información en mapas multidimensionales que son fácilmente distribuibles y accesibles mediante APIs que permiten a los usuarios acceder a los mismos mediante operaciones básicas (get, put, contains, remove). Este enfoque permite la segmentación de los datos para ser leídos y procesados utilizando frameworks basados en MapReduce (p.e. Apache Hadoop). Otro factor clave del éxito de NoSQL en lo que respecta a Big Data es que proporciona diferentes formas o modelos para almacenar los datos, lo cual permite implementar soluciones específicas para cada dominio, por complejo que este pueda ser. Según [18], los principales modelos de datos que se manejan en NoSQL son:

1) Key-Value

Asocia los valores (simples o compuestos) mediante llaves, lo cual facilita su acceso de forma directa. La mayoría de los manejadores de bases de datos que soportan este modelo son derivados de Amazon Dynamo. Entre los más nombrados por diferentes autores, encontramos: Amazon Dynamo, Voldemort [31], Memcached [32], Redis [33] y Riak [34].

2) Document-Oriented

Se asemeja al modelo Key-Value en que utiliza llaves para acceder a los valores almacenados. La diferencia radica en que los valores (documentos) proporcionan alguna estructura o codificación de los datos específica (p.e. XML, JSON, BSON). Los manejadores más nombrados son MongoDB [35] y CouchDB [36].

3) Column-Oriented

Conocido también como Tabled-Based, permite almacenar los datos en tablas tridimensionales indexadas mediante un Row-Key (Similar a Key-Value), que puede ser un hash que identifica una fila. Contiene un Column-Key y un Timestamp por cada dato almacenado. La mayoría de los manejadores de bases de datos que soportan este modelo son derivados de Google BigTable. Entre los más nombrados por los diferentes autores, encontramos: Apache Cassandra [37] y Apache HBase [38].

4) Graph-Oriented

Permite representar las múltiples relaciones entre diferentes entidades y a las entidades en sí, empleando la abstracción y los conceptos matemáticos de grafos para dicha representación. El manejador más nombrado, que soporta este modelo, es Neo4j [39].

Es importante tener en cuenta, cuando se comienza a trabajar con NoSQL, que estas bases de datos -al trabajar en entornos distribuidos- se ven afectadas o implicadas en lo que se conoce como el teorema de CAP (*Consistency, Availability, Partition Tolerance*) [26], el cual manifiesta que sólo se pueden garantizar dos de las tres características que componen sus siglas. Lo anterior hace que los diferentes sistemas manejadores de datos NoSQL ubiquen sus mayores fortalezas (y debilidades) alrededor de estas características. Por ejemplo, apache Cassandra es más fuerte en *A* (Availability) y *P*

(Partition Tolerance), mientras que MongoDB en *C* (Consistency) y *P* (Partition Tolerance).

C. Cloud Computing

Hace unos años, la capacidad de computación de una compañía estaba altamente definida por la cantidad de dinero que esta tuviese para invertir en servidores e infraestructura que permitieran satisfacer la necesidad de procesamiento y, cada vez que esta fuese mayor, se hacía necesario ampliar las características (procesador, memoria, almacenamiento) de estas máquinas o reemplazarlas por otras superiores. A esta técnica se le conoce como escalamiento vertical "Scale Up". A la par de este tipo de escalamiento surge otra técnica conocida como escalamiento horizontal "Scale Out" la cual se fundamenta en dividir la carga trabajo de manera paralela en diferentes nodos computacionales de menor capacidad que los usados en el escalamiento vertical, permitiendo así que compañías puedan tener acceso a arquitecturas de alto rendimiento a un menor costo, pero añadiéndole el grado de complejidad que conlleva administrar este estilo de computación distribuida. Basado en lo anterior surgen preguntas como ¿qué estilo de escalamiento usar?, ¿cuántos nodos de computación configurar?, ¿cuánto dinero invertir para tener una arquitectura de alto rendimiento? Y, si la carga de trabajo aumenta o disminuye considerablemente, ¿qué sucede con la arquitectura? La respuesta a estas y otras preguntas es una nueva tendencia tecnológica conocida como computación en la nube "Cloud Computing". Cloud Computing se encuentra definido por [3] como "*Poderosas arquitecturas de computación basadas en sistemas de virtualización de software que lucen como un computador físico, pero con especificaciones flexibles como número de procesadores, tamaño de memoria, tamaño de disco y sistema operativo*" y por [40] como "*Un estilo de computación en la cual capacidades de escalabilidad y elasticidad son ofrecidas como servicio usando tecnologías de internet*". Hoy en día, juega un papel fundamental en el procesamiento de la información ya que permite a las empresas contar con arquitecturas altamente escalables que se pueden expandir o reducir según la carga de trabajo, sin necesidad de realizar una inversión en hardware costoso, el cual, con el paso del tiempo, tiende a volverse obsoleto.

Relación con Big Data Analytics…

La necesidad de analizar la información surgió desde el mismo momento en que los sistemas transaccionales comenzaron a registrar datos, los cuales, consolidados de diferentes formas, permiten generar información útil para la toma de decisiones. Las técnicas tradicionales que se implementaron en los comienzos de los sistemas OLAP tienen qué ver con la primera generación de Data Warehousing, donde la técnica más usada era el procesamiento por lotes. Esta técnica funciona bien cuando el flujo de entrada de datos se mantiene constante y los resultados del procesamiento pueden ser usados una vez hayan terminado todos los procesos en lote. Sin embargo, cuando tratamos con Big Data, donde los resultados son útiles sí y solo sí, los períodos de tiempo de adquisición y procesamiento de los datos son cortos (tiempo real), generando así la necesidad de tener arquitecturas que puedan procesar, a velocidades altamente escalables, datos de tamaño altamente volátil, en el menor tiempo posible. Estos tiempos de respuesta no se podrían lograr utilizando los métodos tradicionales de procesamiento y es por esto que se dice que Big Data trae consigo un cambio de paradigma en el proceso del análisis. Estos cambios de paradigma han desencadenado un amplio conjunto de tecnologías, basadas en la computación distribuida y el almacenamiento de datos complejos, conocidas como NoSQL.

Según [28], la habilidad de crear valor a partir de Big Data depende de la eficiencia de los procesos, la cual se encuentra dirigida en buena parte hacia la disminución de costos de almacenamiento y computación por unidad de datos (esto debido a que no está garantizado que ese procesamiento arrojará resultados que generen valor, contrario a lo que sucede con las aplicaciones tradicionales de Data Warehousing) y a que los resultados deben estar disponibles rápidamente ya que estos sólo son significativos si son entregados de manera oportuna (o en tiempo real) como es el caso de las recomendaciones y los anuncios publicitarios. En satisfacción de lo anterior, se da la necesidad de plantear arquitecturas de alto rendimiento, altamente escalables, expandibles, elásticas y de bajo costo, siendo allí donde juega un papel muy importante la otra tendencia tecnológica, conocida como computación en la nube o "Cloud Computing". En la actualidad, grandes representantes de la computación en la nube, como es el caso de Amazon Web Services [41], están haciendo una gran apuesta al área del Big Data Analytics poniendo poderosas herramientas de computación en la nube al alcance de otras organizaciones emergentes que se encuentran ubicadas en la ciencia, la industria y el Gobierno. Como todo nuevo paradigma, es de esperar que su masificación tome un tiempo: Según Gartner [42], esta espera será cuestión de 5 años.

VI. ARQUITECTURA GENERAL PARA LA IMPLEMENTACIÓN DE BIG DATA ANALYTICS

Extraer el **Valor** (conocido en este documento como la cuarta "V") ha generado diversos retos computacionales ya que las tecnologías tradicionales utilizadas para el procesamiento de datos no logran satisfacer todas las necesidades referentes a escalabilidad, rendimiento, almacenamiento, tiempo de procesamiento, entre otras. Para abordar estos retos las tecnologías de Big Data se apoyan en tres pilares fundamentales: Sistemas de Archivos Distribuidos, Bases de Datos Escalables y Software de Procesamiento en Paralelo.

En busca de estos pilares, muchas herramientas de Big Data Analytics se han basado en lo que es conocido como el Ecosistema Hadoop [8], el cual tiene como sus principales componentes a HDSF (Sistema de Archivos distribuidos), MapReduce (framework el cual permite dividir y paralelizar los cálculos complejos entre un número indefinido de ordenadores) y Hadoop Common (conjunto de librerías que soportan varios subproyectos de Hadoop). Alrededor de estos componentes se han generado una gran variedad de proyectos se pueden integrar con Hadoop, para conseguir mayor potencia y capacidad de especialización en los proyectos de Big Data. Empresas referentes en el tema presentan como propuesta de valor arquitecturas basadas en dicho ecosistema, las cuales son afinadas según su área de aplicación (Retail, Gobierno, Salud, Telco) y puestas a disposición de los clientes para que estos comiencen a extraer el valor de sus datos sin preocuparse por los aspectos técnicos que esto conlleva.

En la Figura 7 se presenta una abstracción de lo que sería una arquitectura base de un sistema para Big Data Analytics, la cual fue generada después de analizar varias herramientas disponibles hoy en el mercado. Revisando esta figura, de abajo hacia arriba, en primer lugar encontramos "Hadoop", el cual es el componente principal de la arquitectura y sobre el cual se soporta todo el sistema de Big Data Analytics. Un nivel más arriba se encuentra el componente "Middleware", el cual a su vez está conformado por dos componentes "Third Party Tools" y "Customs Tools". A este primero pertenecen todas aquellas aplicaciones disponibles en el mercado que interactúan y/o se integran con Hadoop para potencializar el análisis de Big Data. Diferentes autores y organizaciones que han implementado este tipo de arquitecturas le han atribuido a este componente -junto con el primer nivel mencionado anteriormente- el nombre de "Ecosistema Hadoop". El componente "Custom Tools" (presente también en el nivel superior) se ha constituido como parte de la propuesta de valor de cada uno de los diferentes sistemas desarrollados bajo este tipo de arquitecturas. En la parte superior se encuentra el componente "Interface" en el cual se agrupan herramientas y aplicaciones que permiten la interfaz con los usuarios finales y/o con otros sistemas.

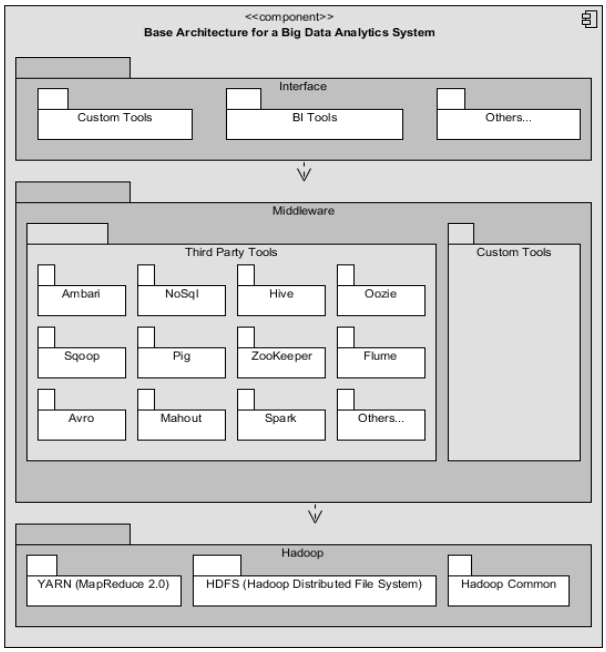

Figura 7. Arquitectura Base para un Sistema de Big Data Analytics

En las siguientes sub secciones se brinda una breve conceptualización sobre algunas herramientas que pertenecen a los diferentes componentes de la arquitectura base:

A. Componente Hadoop

La plataforma de computación en la nube Hadoop es un framework que permite el procesamiento distribuido de grandes cantidades de datos sobre clústeres de computadores a través de un simple modelo de programación y permite a los programadores crear programas de procesamiento en paralelo sin tener experiencia en el desarrollo de sistemas distribuidos. Es ampliamente usado en el cálculo y la administración de Big Data. Sus principales componentes incluyen MapReduce, HDFS, HBase, ZooKeeper, entre otros. La plataforma es altamente eficiente, altamente confiable, de bajo costo, altamente expandible y libre. Muchas aplicaciones basadas en Hadoop son realmente exitosas. Por ejemplo, Facebook lo usa para analizar sus registros diarios y hacer minería de datos, China Mobile lo usa para analizar los datos de sus clientes con el fin de mejorar la configuración de sus redes de comunicaciones [43]. Los tres componentes principales de Hadoop son:

B. MapReduce

Es un modelo de programación y ejecución para el procesamiento y generación de grandes volúmenes de datos, impulsado por Google y desarrollado por Yahoo entre otras compañías. Es basado en el método divide y vencerás, y funciona dividendo recursivamente un problema complejo en muchos sub-problemas hasta que estos sean escalables, de tal manera que puedan ser resueltos directamente. Después, estos sub-problemas son asignados a un clúster de equipos de trabajo y son resueltos por separado y de manera paralela. Finalmente, las soluciones de los sub-problemas son combinadas para brindar una solución al problema original. El método divide y vencerás es implementado a través de dos pasos: Map y Reduce [3]. Según [15], en el análisis de Big Data, Google's MapReduce es el primer gran paradigma para el procesamiento de datos, el cual fue propuesto para facilitar el desarrollo de aplicaciones altamente escalables, tolerantes a fallos y distribuidas a gran escala.

1) HDSF (Hadoop Distributed File Systems)

Es un sistema distribuido de almacenamiento de archivos el cual provee altas capacidades, altas tasas de transferencia, alta tolerancia a fallos, alta expansibilidad y bajo costo de almacenamiento para enormes cantidades de datos. HDFS tiene una arquitectura maestro/esclavo, la cual consiste en un solo NameNode y muchos DataNodes. Este primero es el servidor maestro encargado de: dividir los archivos en uno o muchos bloques los cuales son almacenados en los DataNodes, almacenar el mapeo de los bloques de un archivo, manejar el espacio de nombres del sistema de archivos y de regular el acceso a los archivos por parte de los clientes. Los DataNodes son responsables de atender las peticiones de lectura y escritura de los clientes del sistema de archivos.

2) Hadoop Common

Corresponde a un amplio conjunto de utilidades comunes que soportan los diferentes módulos de Hadoop, sobre las cuales se construye el core de una implementación Hadoop típica.

C. Componente Middleware

Este componente se encuentra conformado por una gran cantidad de utilidades y herramientas que permiten la comunicación entre los componentes “Hadoop” e “Interface”. Básicamente, como su nombre lo indica, permite hacer una mediación entre estos componentes a través de distintos lenguajes, APIs o herramientas, cada una con un propósito específico. Estas utilidades se encuentran divididas en dos componentes principales:

1) Third Party Tools

Este componente hace referencia a todas las utilidades que se encuentran cobijadas por proyectos de diferentes organizaciones (dentro de ellas, una de las más importantes es Apache). Estas utilidades, herramientas y librerías tienen la característica de ser estándares para los propósitos para los cuales fueron construidas. Algunas de las herramientas y tecnologías más importantes de este componente son:

- Ambari [44]. Herramienta para configuración de clústeres Hadoop.
- NoSql. Bases de datos escalables para almacenar y procesar grandes volúmenes de datos [2]. Descrita en la sección V.B.
- Apache Hive [45]. Sistema de data warehouse sobre Hadoop.
- Apache Oozie [46]. Orquestador de tareas relacionadas con el ecosistema Hadoop.
- Apache Sqoop [47]. Herramienta de ETL diseñada para transferir de forma eficiente información entre Hadoop y bases de datos relacionales.
- Apache Pig [48]. Proporciona un lenguaje de alto nivel para simplificar a los usuarios de Hadoop en análisis de grandes volúmenes de datos.
- Apache ZoopKeeper [49]. Herramienta de sincronización de clusters Hadoop.
- Apache Flume [50]. Herramienta para capturar, analizar y monitorizar datos de ficheros de log.
- Apache Avro [51]. Sistema de serialización de datos.
- Apache Mohout [52]. Plataforma de aprendizaje autónomo y data mining construida sobre Hadoop.
- Apache Spark [53]. Altenativa a Hadoop que se basa en el almacenamiento de datos en memoria.

2) Custom Tools

Se encuentra conformado por aquellas herramientas middleware que ofrecen las diferentes empresas a sus clientes para satisfacer sus necesidades específicas de negocio. Dichas herramientas pueden variar dependiendo del mercado objetivo y son parte del valor agregado diferenciador de proveedores de sistemas de Big Data Analytics. Ejemplos de ellas son APIs, aplicaciones y componentes diseñados a la medida para realizar algún tipo de procesamiento específico y/o para servir de puente entre la capa de procesamiento y la de interfaz.

D. Componente Interface

A este grupo pertenecen todas las utilidades y herramientas que permiten que los usuarios y/u otros sistemas interactúen con el sistema de Big Data Analytics implementado. Se encuentra dividido en las siguientes categorías:

1) Custom Tools

También hace parte del valor agregado de las diferentes empresas que ofrecen plataformas de Big Data y está compuesto por aquellas aplicaciones Front-end que permiten a los usuarios del sistema hacer uso de la información analizada para su comprensión, visualización e interpretación.

2) BI Tools

Herramientas de inteligencia de negocios que interactúan con el sistema de Big Analytics para extraer el valor de los datos (la cuarta "V"). Algunos ejemplos de estas herramientas son: Pentaho [54], SAP Business Intelligence [55], Business Intelligence Suite de Microsoft [56], Microsoft Excel [56], Jasper Reports [57], Cognos Business Intelligence de IBM [58] y Oracle Business Intelligence [59].

3) Otros

Otros componentes que de alguna u otra manera interactúan con el sistema y no hayan sido categorizados como *Custom Tools* o *BI Tools*.

VII. OPORTUNIDADES, RETOS Y TENDENCIAS

Hoy en día, las empresas se encuentran más preocupadas por entender patrones, comportamientos y posibles tendencias a partir de los datos que son almacenados por medio de los sistemas transaccionales. A partir de lo anterior, se encuentran diversas oportunidades que permiten llegar a este entendimiento gracias a los conceptos y la revolución que trae Big Data, permitiendo a su vez que se marquen tendencias en las tecnologías, técnicas y herramientas para abordarlas más fácilmente. Sin embargo, esto implica una serie de retos que deben ser tenidos en cuenta para evitar implementaciones que puedan no resultar adecuadas en términos de viabilidad técnica o económica, principalmente, ya que se debe recordar que: "Big Data Analytics es la capacidad de extraer valor de Big Data, sin tener la certeza de que ese valor exista. Esto implica estar expuestos a la volatilidad y el ruido que pueden tener estos datos durante su procesamiento que hacen que dichos datos, en ocasiones, carezcan de valor". En la presente sección se abordarán las oportunidades, retos y tendencias que se presentan en el análisis de Big Data, teniendo en cuenta algunos autores y organizaciones relevantes en la materia:

A. Oportunidades

Es claro que la toma de decisiones a partir de Big Data es, hoy en día, uno de los aspectos que más influencia el desarrollo de los diferentes sectores productivos, en sus respectivas áreas de aplicación (Gobierno, Industria y Academia). Agencias como el Instituto Nacional de Salud (NIH) y la Fundación Nacional de Ciencias (NSF) de Estados Unidos comparten esta afirmación [3]. Otras instituciones a nivel global como la ONU (Organización de las Naciones Unidas) postulan a Big Data como uno de los medios para mejorar la calidad de vida a nivel mundial. Tanto es así que esta organización, en el Panel de las Naciones Unidas sobre la Agenda Post-2015 [60], concluyó que "para poder alcanzar los objetivos del Milenio necesitan una revolución de datos". Para la ONU, Big Data abre numerosas oportunidades para el desarrollo, las cuales son de gran pertinencia. Siendo consecuentes con lo anterior, este tipo de organizaciones se mantiene en investigación para determinar cómo beneficiarse de estas tecnologías y técnicas que ofrece en este caso Big Data, y cómo pueden mejorarse desde el punto de vista de cada área de aplicación. Por lo anterior, se crean iniciativas a nivel nacional y mundial que

permiten disminuir la brecha tecnológica que se crea alrededor de estos nuevos conceptos.

De acuerdo con estudios como el del instituto de McKinsey [61], si Big Data se usa de manera eficaz, se pueden obtener grandes beneficios que van hasta la transformación de una economía y la llegada de una nueva ola de crecimiento productivo para una región. Las ventajas que se pueden obtener a través del aprovechamiento de Big Data (la cuarta V) se ilustran en la Figura 8. Estas incluyen, el incremento de la eficiencia operativa, mantener informada a la dirección estratégica, mejora del servicio al cliente, identificación y desarrollo de nuevos productos y servicios, identificación de nuevos clientes y mercados, oportunidad para manejar el "time-to-market", fácil cumplimiento de normas, entre muchas otras.

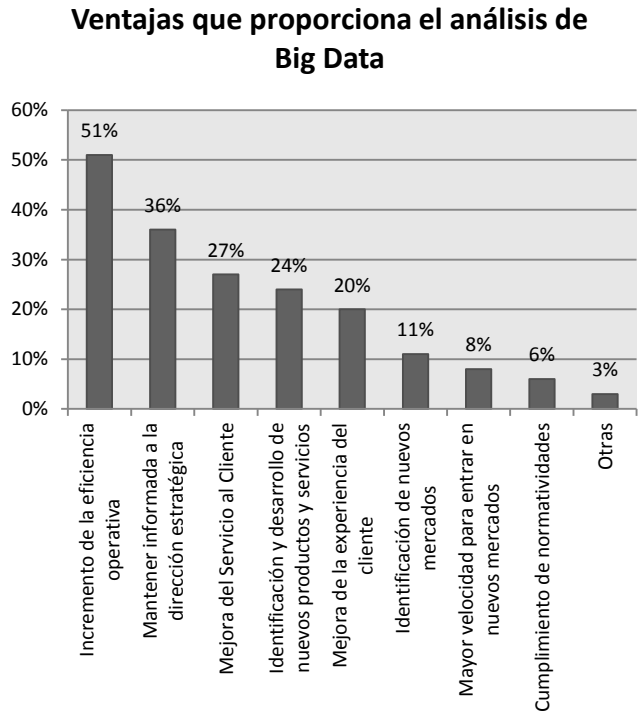

Figura 8: Ventajas que proporciona el análisis de Big Data

Esta figura se encuentra basada en un estudio realizado en [3], el cual se basó en analizar aproximadamente 560 empresas que se encuentran involucradas con Big Data. La mayoría de estas empresas (más del 50%) afirmaron que la mayor ventaja que proporciona el hecho de aprovechar el valor que se obtiene de Big Data es el incremento en la eficiencia de su quehacer, al entender mejor sus procesos. Lo anterior pone a este artículo de regreso al estudio y los planes de la ONU nombrados anteriormente, donde se busca primero "entender el mundo" y luego intentar mejorarlo. Luego entonces, este entendimiento puede ser posible gracias a Big Data Analytics, cuyo insumo principal son los datos que se generan a través de diferentes medios como sensores y redes sociales. La ONU le ha llamado a esto la "revolución de los datos" y, en el artículo referenciado en [9], habla de Big Data Analytics como la capacidad de "ponerle sentido" a los datos que se generan alrededor de varios sectores y la academia.

Entre varios ejemplos citados en [9] y [62], este artículo destaca 11 que ilustran claramente cómo es posible beneficiarse de "la revolución de los datos" y el valor que puede extraerse de los mismos:

1) Entender procesos demográficos y migratorios

Por ejemplo, rastreando teléfonos celulares vía GPS se ha logrado comprender patrones de migración y de formación de grupos sociales en ciudades. También se pueden identificar tendencias de desempleo mucho antes que los reportes oficiales, a través del análisis de redes sociales. Esto permite suministrar oportunidades a las personas o establecer planes de choque ante variaciones o patrones que se encuentren por fuera de los límites de control establecidos.

2) Identificar hábitos y problemas sociales

Gracias al análisis de redes sociales como Twitter y Facebook, se ha podido detectar enfermedades, hábitos y problemas que muchas veces las personas no comparten con sus médicos de cabecera, pero sí son divulgadas en estos medios. Esto permite identificar las tendencias o patrones frente a hábitos y problemas como el consumo de drogas o alcohol.

3) Mejorar los sistemas que alertan desastres

Hace algunos años, se descubrió que era posible localizar terremotos a través de los mensajes registrados en Twitter. Ahora, el Servicio Geológico de los Estados Unidos está obteniendo un 90% de exactitud en dicha localización a través del análisis del incremento de mensajes y tendencias de Twitter sobre actividades sísmicas. En otros casos, se ha logrado analizar los datos arrojados por sensores que monitorean la actividad oceánica para predecir tsunamis y anticipar sus riesgos de manera efectiva.

4) Comprender tendencias económicas

Sobre este tópico existen diversos avances ya que es uno de los más estudiados. Por ejemplo, Walmart utiliza técnicas de Big Data Analytics para predecir posibles comportamientos de sus clientes y, de esta manera, ofrecer promociones y crear productos y servicios que apuntan a dichos comportamientos. Otro ejemplo reciente sobre este tópico es un sistema, desarrollado por investigadores del MIT, capaz de recoger datos diarios sobre precios de bienes vendidos o promocionados en la web para luego estimar la inflación del país con mayor precisión y rapidez que los métodos tradicionales.

5) Detectar riesgos de epidemias y pandemias en tiempo real

Gracias al análisis de Big Data se puede inclusive detectar la posibilidad de brotes epidémicos en cualquier momento, por ejemplo, monitoreando las búsquedas que los usuarios hacen en internet sobre síntomas típicos de algunas enfermedades, en lugares específicos. Dos ejemplos importantes de esto son: Google Flu Trends [63] y Google Dengue Trends [64], capaces de monitorear la evolución de la gripe y el dengue, principalmente.

6) Descubrir cambios topográficos, patrones de tráfico y emisiones – Ciudades Inteligentes

El entendimiento de los patrones que se generan en una ciudad como el tráfico, las emisiones y los cambios topográficos, permitirá la evolución de las ciudades en lo que se conoce como "Ciudades Inteligentes", lo que está constituido como una tendencia en el mundo. Proyectos como "Ciudad Creativa Digital" [65], busca ser la base de las ciudades inteligentes, por medio de la instalación de sensores capaces de transferir datos en tiempo real sobre la actividad de cualquier ciudad, permitiendo varios usos como: controlar semáforos, organizar el tráfico y potenciar sectores de la ciudad como el turismo.

7) Entender el cambio climático

Con este tópico se busca, entre otros, optimizar el aprovechamiento del clima para proyectos agroindustriales y de infraestructura, prevenir desastres naturales, permitir el ahorro de recursos como el agua y alertar efectivamente sobre fenómenos como el calentamiento global y sus efectos. Existen muchas aplicaciones desarrolladas en centros de investigación que utilizan datos provenientes de sensores (de humedad, temperatura, presión, etc.) para encontrar variaciones, patrones y tendencias que apoyen la toma de decisiones.

8) Mejorar los servicios públicos

Para mejorar la calidad de vida de las personas, sin lugar a dudas, un punto clave son los servicios públicos. Hacer que su suministro sea más inteligente, de mejor calidad y menos costoso es la gran oportunidad que puede brindar el uso eficiente de Big Data. Por ejemplo, se puede monitorear un servicio en busca de información útil que sirva para mejorarlo (Suministro de agua, energía, telecomunicaciones, servicios de salud). Un ejemplo de lo anterior es una aplicación llamada "Ubidots" [66], que monitorea las condiciones de higiene de 25 hospitales en América Latina. Esta aplicación es capaz, entre otras cosas, de monitorear la actividad de los equipos de cada hospital y conocer la tasa de ocupación del hospital, lo que permite conocer la situación real del mismo, en tiempo real.

9) Fortalecer el trabajo colaborativo

Este tópico aplica en diversos campos de acción y busca mejorar la calidad de vida desde cada punto de vista, a través de la colaboración a gran escala como eje principal. Dos ejemplos de lo anterior son: *1)* CoCoRaHS [67], la cual es una red de voluntarios que miden precipitaciones pluviales. Con la información que se recoge de esta plataforma, las comunidades locales pueden controlar las epidemias de mosquito y mejorar la panificación urbana. *2)* IPython [68] es una plataforma diseñada como apoyo para la publicación de artículos científicos, de forma colaborativa.

10) Salvar vidas

El análisis de Big Data tiene un gran campo de acción en la salud y el cuidado humano. Por ejemplo, según [62], por medio de la recolección y el análisis de millones de datos de las unidades de cuidados intensivos de neonatos se puede alertar de forma temprana señales de posibles infecciones.

11) Mantener el equilibrio entre un gobierno y sus ciudadanos

Dado que existen medios tan poderosos como las redes sociales, blogs, sitios que permiten registrar comentarios, etc., un gobierno puede vigilar sus acciones (por ejemplo, legislativas) a través del análisis de estos datos para tomar decisiones que permitan mantener el equilibrio de una ciudad, un país o toda una región.

B. Retos

Las oportunidades por lo general traen consigo retos. Mientras en la sección anterior (sección A) se describen algunas de las oportunidades de Big Data Analytics, que pueden sonar muy atractivas e interesantes, por otro lado aparece una cantidad significativa de retos o desafíos que deben abordarse o, por lo menos, tenerse en cuenta antes de extraer los beneficios que ofrecen tales oportunidades [3]. En la definición de Big Data suministrada por Gartner en [40] "Big data son aquellos activos de información de gran volumen, variedad y velocidad que demandan formas de procesar la información, innovadoras y efectivas en costo, para mejorar el entendimiento y la toma de decisiones", se presentan dos aspectos de los cuales se desprenden los principales desafíos de Big Data Analytics: *Innovación y efectividad en costos*, que corresponden a los medios por los cuales será logrado el valor esperado a partir de los datos. Estos aspectos inherentes en las definiciones de Big Data, junto con las oportunidades que se vislumbran, desencadenan una serie de retos relacionados con los datos, tales como: Captura, Almacenamiento, Transmisión, Procesamiento, Curación, Análisis, Visualización, Seguridad, Escalabilidad, Desempeño y Consistencia; y estos a su vez, son tangenciales a los retos que enfrentan las organizaciones para ingresar y tener éxito en el mundo de Big Data Analytics. A continuación, se describen brevemente algunos de los desafíos más relevantes, en los cuales se centra gran parte de las investigaciones académicas y gubernamentales, y los productos y servicios ofrecidos por la industria, en la actualidad:

1) Retos relacionados con la captura y el almacenamiento de los datos

El primer reto a abordar tiene qué ver con la accesibilidad de los datos, dado que Big Data Analytics implica más I/O, teniendo que lidiar con un fenómeno que ha existido por décadas y tiene qué ver con que los avances en I/O no se encuentran equilibrados con los avances en CPU, contando éstos últimos con una ventaja de una década sobre los primeros [3]. Otros retos a abordar son: La distribución (alta concurrencia y el manejo de un throughput por cada servidor), replicación, migración, desempeño, confiabilidad y escalabilidad.

2) Retos relacionados con la transmisión de los datos

Cada vez es más frecuente el uso de la computación en nube (cloud computing) y los esquemas de distribución de datos y tareas en Big Data Analytics, con el fin de responder a las exigencias esto implica. A la hora de mover datos de un sitio a otro para su procesamiento, surgen varios desafíos que hacen que dicha transmisión no sea un proceso trivial. La transmisión de Big Data en las tecnologías orientadas a la nube y a sistemas distribuidos tiene diferentes retos a abordar, entre

los que se destacan: El ancho de banda, la seguridad y la integridad de los datos [3].

3) Retos relacionados con el procesamiento y curación de los datos

El procesamiento y curación de datos comprende desde el descubrimiento de los datos hasta su recuperación, aseguramiento de calidad, adición de valor, reutilización y preservación en el tiempo [2]. El inconveniente con las herramientas tradicionales (principalmente los modelos que trabajan con datos estructurados) es que éstas no tienen la capacidad de manejar estos procesos con Big Data de forma eficiente, teniendo en cuenta las definiciones relacionadas con las 3V, discutidas en la sección II. Lo anterior hace que el análisis de Big Data deba incluir técnicas innovadoras desde la captura, representación y almacenamiento de los datos hasta su visualización después de los procesos de análisis, teniendo en cuenta que dichas técnicas deben poderse aplicar a un bajo costo.

4) Retos relacionados con la seguridad de los datos

Cuando se habla de almacenamiento y procesamiento paralelo distribuido de datos a gran escala, surge uno de los principales atributos de calidad a tener en cuenta: La seguridad de la Información. Dado que los entornos en los cuales se desenvuelven la mayor parte de técnicas y tecnologías de Big Data son cloud-enabled o manejan arquitecturas de datos compartidos la mayoría de las veces, las organizaciones pudieran tener la tendencia a desconfiar de dichas arquitecturas y entornos que se habilitan para el análisis de Big Data, lo que podría ser una de las causas principales de la brecha que existe entre las organizaciones y este tipo de tecnologías. La seguridad de los datos en análisis de Big Data tiene dos preocupaciones principales, según [9]: *(1)* La *Privacidad*, que es la más delicada con implicaciones conceptuales, legales y tecnológicas, y se encuentra definida por la Unión Internacional de Telecomunicaciones (ITU) [69] como "el derecho de las personas de controlar o influenciar sobre qué tipo de información relacionada con ellas puede ser revelada". En este caso, sus implicaciones deben tenerse en cuenta en el uso de Big Data como apalancador de desarrollo durante la adquisición, almacenamiento, retención y presentación de los datos. (2) El *Acceso* y *Participación*, que consiste en que cualquier persona pueda acceder a datos valiosos, incluso cuando estos se encuentran almacenados en organizaciones privadas. Los retos en este caso son numerosos y tienen que ver principalmente con aspectos legales, reputación, seguridad nacional, competitividad, secretos de industria y la falta de incentivos y estructuras de información que pueda ser pública. Ejemplos pertinentes de este tipo de información son: La Historia Clínica, el Perfil Profesional (Curriculum) y el Perfil Académico (Historia Académica). Este tipo de información debe ser pública y sin costos de divulgación. Obviamente, el uso malintencionado que pudiera darse a esta información es un tema de gran profundidad, por lo que el tratamiento de este riesgo constituye el mayor de los retos relacionados con la seguridad.

5) Retos relacionados con el análisis de los datos

Este reto parte de la pregunta "*¿Qué nos están diciendo realmente los datos?*". Dependiendo del tipo de análisis que sea llevado a cabo y el tipo de decisiones que se tomarán con estos datos, los retos serán diferentes. Sin embargo, según [3] y [9], existe un factor común inherente a los retos que tienen qué ver con el análisis de los datos, el cual se encuentra dividido en tres aspectos: *(1) Obtener una población correcta de los datos*, que consiste en la obtención de datos limpios (libres de ruido) y veraces (que puedan ser verificables), con el fin de evitar hechos falsos que puedan alterar la percepción de la realidad. *(2) Interpretar los datos*, que consiste en emitir conceptos, realizar proyecciones y tendencias a partir de los datos recolectados. Este reto tiene qué ver en gran parte con el anterior, ya que sin los datos adecuados, las interpretaciones posiblemente serán incorrectas. *(3) Definir y detectar anomalías*, que consiste en que es realmente complejo caracterizar las anomalías en ecosistemas humanos para futuros análisis.

6) Retos relacionados con la arquitectura

Dado que Big Data se encuentra soportado en su mayor parte por Cloud Computing, diferentes autores plantean una serie de inquietudes, como las mencionadas en [16], que implican retos arquitectónicos que deben ser abordados por los implementadores de sistemas de análisis de Big Data. Estos retos consisten principalmente en: *(1)* Determinar cómo la integración de servicios en la nube permite el manejo de Big Data; *(2)* Cómo facilitar el escalamiento de sistemas legados a través de Cloud Computing y las técnicas y tecnologías relacionadas con Big Data; y *(3)* Cómo los servicios Cloud-enabled (XaaS) manejan las grandes necesidades y desafíos que plantea Big Data: Las 3 Vs. Finalmente, estos planteamientos llevan a la necesidad de implementar Frameworks genéricos de gestión de workflows de Big Data con el objetivo de implementar sistemas flexibles y reutilizables en las diferentes organizaciones, en forma de servicios (XaaS, donde X pueden ser estos Frameworks).

7) Retos relacionados con la visualización de los datos

El principal objetivo de la visualización de los datos es la representación del conocimiento para el entendimiento del ser humano, utilizando diferentes mecanismos que resultan ser aproximaciones más intuitivas que sofisticadas. Según [3], en el campo de Big Data, el reto principal frente a esta representación se encuentra en el gran tamaño y diversidad de los datos. Actualmente, la mayoría de las herramientas de visualización deben enfrentarse a inconvenientes de desempeño y escalabilidad. Esto hace que las técnicas y tecnologías que se vienen utilizando por años deban re-pensarse para abordar estos retos de una forma más efectiva. Esto último -de hecho- se ha venido desarrollando en organizaciones como Microsoft [70], JasperSoft [71], Pentaho [72] y Tableau [73], por ejemplo.

8) Retos relacionados con las Organizaciones

La revolución de los datos masivos no es una tendencia: siempre ha existido pero no se había explotado hasta ahora que sale a la luz el conjunto de conceptos que encierran el mundo de Big Data. Más que las organizaciones ser un reto para Big Data, es la Big Data un gran reto para las organizaciones, ya que, a pesar de las múltiples oportunidades que se vislumbran con esta, las empresas aún están lejos de aprovecharla al 100%. El verdadero reto no está tanto en crear las arquitecturas que permitan implementar Big Data (y su posterior análisis) en las

organizaciones si no en ser capaces de separar, en el menor tiempo de respuesta posible, lo que es relevante y lo que no de todo el volumen de datos se generan. Es por esto que el foco está puesto en entender lo que dicen los datos y cómo las organizaciones sacan partido al valor que puede extraerse de ellos. Para lo anterior, la mayoría de los autores plantean que las organizaciones deben primero que todo determinar los objetivos de análisis requeridos para luego decidir qué datos almacenar y cómo regresarán en forma de información valiosa. Las organizaciones que no se planteen objetivos claros desde el principio, simplemente almacenarán datos que no sabrán cómo aprovechar en un futuro, resultando de este ejercicio posibles frustraciones y el abandono de estas tecnologías. Por otro lado, las organizaciones tendrán que incrementar sus esfuerzos para lidiar con todo el impacto cultural que trae Big Data con respecto a la privacidad de los datos y el recelo de los usuarios para ofrecerlos.

C. Tendencias

Cuando se habla de tendencias en Big Data, la mayoría de autores se refieren a tres grupos principales: Tendencias en Almacenamiento, Tendencias en Comunicaciones y Tendencias en Software Stack. Este artículo se encuentra enfocado en las tendencias que existen en cuanto al Software Stack, dentro del cual están los sistemas, aplicaciones, APIs, arquitecturas de software, servicios de integración B2B y B2C, appliances, distribuciones, entre otros componentes.

Según [74], la orientación de las empresas hacia los datos está generando data warehouses cada vez más capaces con el pasar de los días. Esto se da porque alrededor del 90% de los datos que se están generando no son estructurados (provenientes de sensores, por ejemplo), lo cual está haciendo que las herramientas se implementen ahora con el objetivo de hacer una mejor extracción de estos datos que cuentan con altos niveles de volatilidad, poca precisión y diversidad de formatos. Entonces, para optimizar o hacer que los data warehouses sean más capaces, se crean aproximaciones que hacen que este tipo de soluciones sean más flexibles y simples, sobretodo permitiendo su uso inmediato sin tener que alcanzar grandes curvas de implementación y adecuación (de allí se hacen populares los términos "plug and play" y "sandbox" para este tipo de tecnologías [74]). Por lo general, estas soluciones son contratadas o adquiridas con pocos fabricantes o, inclusive, con uno solo por la facilidad de implementación que esto ofrece. Según Gartner y su reporte de "cuadrantes para Sistemas de Data Warehousing" [22], las empresas líderes en sistemas que proporcionan soluciones de data warehouse y análisis de datos entienden esta tendencia y la proyectan para proporcionar soluciones de este tipo, que permiten a las compañías concentrarse más que todo en sus necesidades, dejando en un segundo plano los esfuerzos técnicos (sin que la planeación e implementación de estas adquisiciones sea una labor despreciable en TI). En la Figura 9, se puede ver el reporte gráfico de los cuadrantes de Gartner. En esta figura se puede observar cómo grandes empresas como Oracle [75], SAP [76], Microsoft [77] e IBM [78] no se han estancado en sus clásicos modelos de análisis de datos. Por el contrario, estas empresas se han mantenido en investigación sobre las principales oportunidades que llegan año tras año y han comprendido claramente las tendencias que ofrecen diferentes tecnologías emergentes como la del presente artículo. Estas empresas han aprovechado el potencial de Big Data -y su posición de líderes en el mercado- para proponer soluciones altamente competitivas, que se han implementado en numerosos y reconocidos clientes. Dentro de este grupo de líderes se destaca a Teradata [11] como la organización número uno en data warehousing, con productos de vanguardia que han logrado superar las expectativas de sus clientes gracias a su integración con Frameworks de distribución de datos como Hadoop [8], Bases de Datos NoSQL, Appliances "sand-box" para cumplir tareas específicas de análisis de datos y su capacidad para ofrecer sus análisis de Big Data como servicio con grandes estándares de calidad.

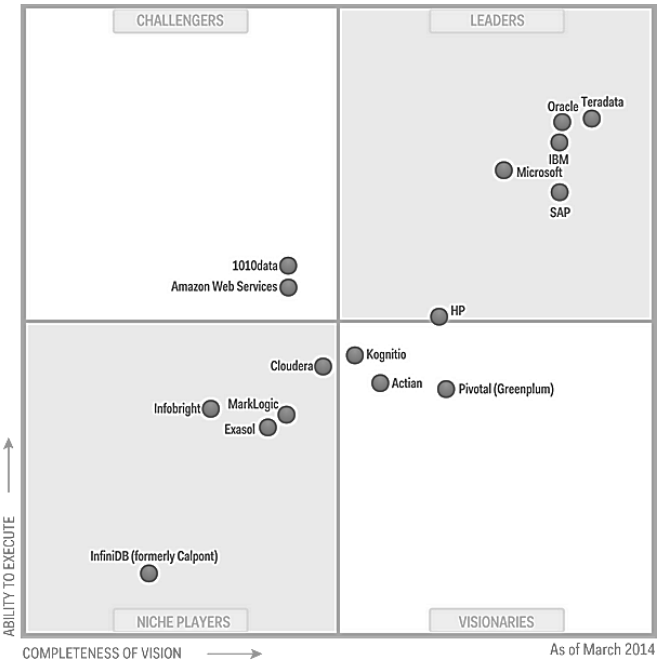

Figura 9: Cuadrantes de Gartner - Data Warehouse 2014

Sin embargo, estos líderes tienen una desventaja primordial: El costo de sus implementaciones continúa siendo elevado, manteniéndolas alejadas de mercados más pequeños (PYMES, por ejemplo) pero que en conjunto pueden representar una buena parte de la población. Es por lo anterior que están tomando cada vez más fuerza alternativas como 1010data [79], Amazon Web Services [80] con sus servicios de Data warehousing (Redshift) y almacenamiento de Big Data (DynamoDB), Cloudera [81], MarkLogic [82], Kognitio [83], Actian [84] y Pivotal [85]. Estas organizaciones tienen dos puntos en común, fundamentalmente: *1)* La implementación de frameworks como Hadoop para facilitar el análisis de Big Data. *2)* Ofrecen sus soluciones de análisis de datos como servicios, por medio de arquitecturas altamente flexibles y escalables.

De las posibles tendencias para Big Data Analytics, el presente artículo hace mención especial sobre las siguientes, consideradas las más importantes por sus autores:

1) Apache Hadoop

Descrito ampliamente en la sección VI. Este Framework ha sido bastante utilizado en la industria y la academia para una gran variedad de tareas, dentro de las cuales, la minería de

logs y el análisis de datos estructurados y no estructurados son los casos de uso más frecuentes [86]. Hadoop optimiza los ambientes de data warehousing por medio de la aceleración de la transformación de los datos [74]. Un ejemplo de lo anterior es convertir flujos de datos no estructurados generados a partir de sensores o redes sociales en datos semiestructurados disponibles para tareas de reconocimiento de patrones y análisis dentro de un data warehouse. El costo es uno de los grandes beneficios de manejar clusters de Hadoop para el procesamiento de datos. La principal tendencia que se puede observar con respecto a la implementación de Hadoop para análisis de Big Data es la orquestación de este Framework -y sus diferentes componentes relacionados- en arquitecturas para el análisis de datos de cualquier tipo de industria (o al menos las más representativas). Estas arquitecturas se encuentran apoyadas por la computación en la nube, haciendo posible la habilitación de estas plataformas de análisis de datos como servicios.

2) NoSQL y Sistemas Híbridos

En el presente artículo se habla de NoSQL como uno de los conceptos clave alrededor de Big Data Analytics (Sección V) ya que los principales retos que trajo consigo Big Data tienen qué ver con las 3 V (Sección II) y una de las principales soluciones a esto se encuentra en la implementación de los motores de bases de datos NoSQL. NoSQL continuará siendo tendencia en la medida que se generen datos no estructurados a grandes velocidades. Es común que hoy en día se encuentren tendencias que apunten a sistemas híbridos entre SQL y NoSQL para tomar lo mejor de cada sistema manejador (ACID y BASE) [2], como es el caso de HadoopDB [87], el cual es una combinación de Hadoop y PostgreSQL [88] teniendo a PostgreSQL como capa de datos, Hadoop como capa de comunicación y almacenamiento, y Hive como capa de traducción de SQL a MapReduce. Existen otras arquitecturas implementadas generalmente como sistemas RDBMS paralelos capaces de conectarse con Hadoop para la carga de datos y la ejecución de tareas MapReduce. La mayoría de estas soluciones ofrecen una especie de semántica MapReduce-SQL nativo. Las tres representaciones más destacadas de este estilo arquitectónico son Pivotal Greenplum [85], Aster Data – Teradata [11] y HP Vertica [89].

3) Data Analytics as a Service (DAaaS)

Según [19], DAaaS es una plataforma analítica extensible, la cual es suministrada utilizando un modelo basado en la nube, donde se encuentran disponibles varias herramientas configurables para análisis de datos. La idea con este tipo de plataformas es que las organizaciones cliente las alimenten con datos empresariales (producto de sus sistemas transaccionales, por ejemplo) y obtengan información concreta y útil para la toma de decisiones. Esta información es generada por aplicaciones analíticas, las cuales orquestan flujos de trabajo específicos para análisis de datos, los cuales son construidos utilizando colecciones de servicios que implementan algoritmos analíticos. Una característica de estas plataformas es que son extensibles, lo que permite manejar diferentes casos de uso o áreas de aplicación (Por ejemplo, Retail, Telecomunicaciones, Salud, Gobierno). El principal beneficio que se obtiene de DAaaS es la reducción de la barrera que existe para la entrada a capacidades analíticas avanzadas por parte de organizaciones que apenas se encuentran introduciéndose en el mundo de Big Data Analytics. Esto permite a estas organizaciones concentrarse más en sus KPI (el qué) que en la forma como se obtendrán (el cómo). Todo lo anterior, a un bajo costo, por tratarse de plataformas habilitadas en la nube (o cloud-enabled).

4) Compresión de Datos

A pesar de que los costos de almacenamiento de datos se han venido reduciendo, el enorme crecimiento en el volumen de los datos que se da con la explosión de datos -de la que se habla al principio del presente artículo- hace que el almacenamiento sea uno de los elementos de costo más grandes dentro de los presupuestos de los departamentos TI [74]. Las tecnologías actuales de compresión de datos utilizan una combinación de métodos orientados a filas y a columnas que permiten almacenar datos para ahorrar espacio y mejorar el desempeño. Estas tecnologías permiten generar 10x valor (Cuarta V) sin generar 10x de sobrecostos.

5) In-database Analytics

Este término hace referencia a las técnicas de análisis de datos que son aplicadas directamente en los DBMS. Esto permite eliminar la necesidad de mover datos entre servidores, optimizando el data warehousing y reduciendo costos de implementación [74]. El análisis de datos dentro de la base de datos incluye una variedad de técnicas para la búsqueda de patrones en grandes volúmenes de datos. Dentro de estas técnicas se incluyen algoritmos de minería de datos implementados dentro de la base de datos, funciones SQL para estadísticas básicas y la integración con lenguajes de programación para estadísticas más avanzadas tales como el Sistema R [21]. El hecho de no tener que mover los datos hacia otras fuentes de almacenamiento para su análisis, permite a los analistas obtener información valiosa en mejores tiempos a costos más bajos relativamente. Adicionalmente, esto permite apuntar hacia atributos de calidad como la seguridad, escalabilidad y desempeño. Algunos ejemplos de minería de datos dentro de las bases de datos incluyen análisis de canasta mercado, identificación de fraude y la predicción de las fluctuaciones de mercado. Las principales compañías que suministran soluciones de data warehousing incluyen análisis "In-database" como una de sus alternativas. Algunas de las más relevantes son: Teradata [11], Oracle [75], IBM Netezza [90], Pivotal Greenplum [91], Sybase [92], ParAccel (Actian) [84], SAS [93] y Exasol [94].

6) Arquitecturas caracterizadas por su temporalidad

La temporalidad que se requiere dentro del análisis de Big Data juega un papel de alto impacto dentro del diseño arquitectónico. Los límites de tiempo y las restricciones pueden requerir aproximaciones costosas y especializadas para determinar la forma como es almacenada y organizada la

información, de tal forma que permitan cumplir con las necesidades puntuales de temporalidad [86]. Por ejemplo, no es lo mismo diseñar una arquitectura para el análisis de audio, video o logs de una aplicación, que diseñar una arquitectura para la elaboración de consultas ad-hoc sobre las ventas o el comportamiento de clientes en una organización tipo Retail. Es por lo anterior que deben existir arquitecturas diferentes según la temporalidad del análisis requerido. En [86] se perfilan las tres principales arquitecturas para el análisis de Big Data, dependiendo de la temporalidad. Estas arquitecturas son, básicamente:

- *Plataformas para Análisis Batch:* Son las plataformas más antiguas y las que tienen la mayor cantidad de implementaciones en el mundo. Sin embargo, con la llegada de Big Data, se pueden ver implementaciones de análisis en Batch con el framework Hadoop. Este modo de operación es útil para minería de grandes volúmenes de datos y tareas Machine Learning. La principal diferencia entre las implementaciones Batch tradicionales y las actuales es que las tradicionales operan en mainframes y las actuales distribuyen las tareas en decenas, cientos y hasta miles de nodos que pueden estar soportados en plataformas en la nube. Lo anterior se traduce principalmente en la reducción de costos. Ejemplos de algunas herramientas que realizan análisis batch son: Apache Mahout [95], GraphLab [96] y Weka [97].

- *Plataformas para Análisis Interactivo:* Es una forma de KDD conducido, de forma interactiva, por un humano mientras explora los datos en tiempo real. Este tipo de análisis se soporta mediante herramientas de visualización, queries ad-hoc, análisis "qué pasa si" y herramientas que permiten realizar agregación sobre grandes conjuntos de datos. La forma tradicional de estas plataformas consiste en herramientas de BI conectadas a data marts relacionales. Sin embargo, existe la tendencia de combinar estas herramientas BI con bases de datos MPP (Masive Parallel Processing), dentro de las cuales se incluyen implementaciones con bases de datos NoSQL. Ejemplos de algunas de estas herramientas son: SAP Business Objects [76], Google BigQuery [98], Apache Drill [99], Google Dremel [100], Apache Pig [48] y Apache Hive [45].

- *Plataformas para Análisis de Datos de Streaming:* Se caracteriza por llevar a cabo el proceso de KDD de forma continua y a grandes velocidades. Ejemplos de este tipo de datos son: Tráfico de redes de computadores, logs, llamadas telefónicas, transacciones ATM, búsquedas web y datos de sensores. Entre algunas de las más importantes implementaciones de estas plataformas se encuentran, Apache Kafka [101] (implementado en LinkedIn), Apache S4 [102] (Implementado en Yahoo), Apache Flume [50], Apache Storm [103] y Apache Spark Streaming [53].

7) Arquitecturas caracterizadas por el almacenamiento y representación de los datos

El almacenamiento y representación física de los datos es un factor a considerar para el análisis de Big Data, en contraste con el principio de independencia de los datos que formuló Codd en la presentación de su modelo relacional [20]. Con la entrada de Big Data, es necesario dar más importancia a este tópico, ya que una buena aproximación puede representar reducción de costos y aumento del desempeño, entre otros atributos de calidad. Las principales tendencias en cuanto a estas arquitecturas son:

- *Bases de Datos distribuidas:* Las tres principales arquitecturas que se discuten en [86] -Hadoop, Bases de Datos para el Procesamiento paralelo masivo (MPP) y los sistemas de computación de alto desempeño (HPC)- se caracterizan por emplear algún modelo de procesamiento paralelo para el análisis de datos. La reducción de costos en Hardware, la comercialización de redes de comunicación de alta velocidad y el incremento de la velocidad de procesamiento han hecho posible estas arquitecturas, las cuales no eran posibles antes principalmente por temas de costos [86]. La distribución de las bases de datos es una tendencia que se seguirá observando junto con otras que tiene qué ver con la forma cómo se comparte el almacenamiento (shared nothing o shared everything) y cómo se abordan los tradeoffs del teorema de CAP [2] (consistencia, disponibilidad y particionamiento). Existe una gran cantidad de Bases de Datos distribuidas. Entre las más representativas se pueden encontrar algunas relacionales como Sybase [92], Oracle [75], Microsoft SQL Server [104] y la mayoría de los DBMS NoSQL que tienen la propiedad de ser tolerantes al particionamiento descritas en la sección V.B, como: Apache Cassandra [37], HBase [38], Memcached [32], Voldemort [31] y Riak [34].

- *Bases de Datos "In-memory":* Utilizan, la mayoría de las veces, la memoria principal de la máquina para el almacenamiento de los datos. Son mucho más rápidas que las bases de datos "disk-oriented". Actualmente son utilizadas tanto para sistemas transaccionales como para sistemas analíticos interactivos y de streaming, donde la latencia y el tiempo de respuesta son críticos. Estas bases de datos se implementan frecuentemente como modelos relacionales sin logging o como modelos llave valor en tablas o mapas hash. Este tipo de almacenamiento de datos incrementan el poder y la velocidad de los data warehouses [74]. Es común hoy en día encontrar representaciones columnares y orientadas a grafos en estas implementaciones. Algunos ejemplos de herramientas que implementan esta arquitectura son: Oracle Database In-memory [75], SAP Hana [76], 1010data [79], Exasol [94] y Kognitio [83].

- *Linked Data Oriented (LOA):* Es un modelo de representación de datos distribuidos a través de

colecciones de links, los cuales son accesibles vía URI (Identificador de Recursos Uniforme). Estas arquitecturas se basan en los principios de publicación en la web descritos en [105], y facilitan el proceso de KDD mediante el descubrimiento de relaciones entre nodos. Esta forma de almacenamiento y organización de los datos, aunque no ha sido definida formalmente como las anteriores, será bastante utilizada los próximos años por la facilidad con la que el conocimiento puede ser representado y accedido. Algunas implementaciones recientes que incorporan esta arquitectura son: Knowledge Graph de Google [106], DBPedia [107], Freebase [108] y Classora [109].

8) Arquitecturas caracterizadas por la plataforma para el cómputo de los datos

Con el fin de explotar mejor la Big Data, han surgido arquitecturas o plataformas que apoyan el procesamiento de los datos, las cuales apuntan a resolver las 3 V, haciendo su principal énfasis en el volumen, en su mayoría. Dentro de estas plataformas o arquitecturas se pueden encontrar como tendencia cuatro grupos principales:

- *Computación Granular (Granular Computing):* Se encuentra basada en el manejo eficiente de Big Data a través de la utilización de gránulos tales como clases, clusters, subsets, grupos e intervalos para separar los datos, con el fin de reducir el volumen de los mismos dentro de diferentes grados de granularidad para aplicar distintos algoritmos de minería de datos según dicha separación [3]. Esta aproximación facilita el proceso de KDD en algunos casos relacionados con Big Data donde el grado de precisión de los resultados no es crítico.

- *Computación en la Nube (Cloud Computing):* Es una de las tendencias más comunes en Big Data y se encuentra descrita ampliamente en la sección V.C. Los tres ejemplos más representativos de esta aproximación son: Amazon Web Services [41], Google AppEngine [110] y Microsoft Azure [77]. El beneficio más importante que ha brindado esta forma de computación es la reducción de costos, permitiendo a las empresas mayores capacidades de procesamiento sin ampliaciones relevantes en los presupuestos de TI (algo que hasta hace unos años era imposible).

- *Computación Bio-inspirada (Bio-inspired Computing):* Consolida técnicas y modelos que se han venido estudiando por varios años sobre la forma cómo el cerebro humano y, en general, la naturaleza, almacena, organiza y procesa los datos. Estos modelos son más apropiados para Big Data ya que tienen mecanismos más eficientes para organizar, acceder y procesar datos que los modelos tradicionales (e inclusive otros que son de vanguardia) [3]. Pueden ser utilizados tanto para el diseño de software como de hardware y comienzan a ser una tendencia aunque por ahora experimental y con costos muy elevados como para pensar en su comercialización.

- *Computación Cuántica (Quantum Computing):* Se basa en la utilización de computadores y componentes cuánticos dentro de sus arquitecturas, capaces de incrementar exponencialmente la capacidad de memoria y procesamiento de los computadores tradicionales [3]. Esto resultaría bastante útil para el análisis de Big Data por el desempeño que se podría obtener en tiempos insuperables por otras tecnologías. Sin embargo, tiene la desventaja de ser extremadamente costoso por ahora, lo que imposibilita su comercialización en masa y hace que actualmente sólo se utilice en Universidades y el Gobierno [3], aunque cabe destacar que éstos están haciendo esfuerzos para masificar su uso.

VIII. TRABAJO FUTURO

Por medio de este artículo, diferentes investigadores y analistas cuentan con un documento base, el cual puede ser utilizado para comprender el enfoque principal de Big Data Analytics y las áreas de aplicación, oportunidades, retos y tendencias, con el fin de generar nuevos trabajos que permitan profundizar y/o aplicar estos temas. Inicialmente, y teniendo en cuenta la pertinencia de esta área de estudio, se proponen los siguientes trabajos que pueden abordarse teniendo en cuenta este artículo:

- Desarrollar un Framework que permita identificar problemas que pueden resolverse con Big Data Analytics dentro de las organizaciones, eliminando ambigüedades en los conceptos y adoptando enfoques generales. Este Framework puede ser lanzado como un modelo que puede empalmarse con procesos técnicos, como la Ingeniería de requisitos, y transversales, como la Gerencia de Proyectos. Con este Framework se busca asegurar el éxito de las implementaciones de Big Data Analytics, partiendo de las necesidades organizacionales o requisitos de negocio.

- Diseñar una Arquitectura de Referencia para la implementación de Big Data Analytics como servicio, basados en la arquitectura general descrita en el artículo.

- Basados en los trabajos anteriores, llevar a cabo un caso de estudio ubicado dentro de alguna de las áreas de aplicación referenciadas en el artículo, desde su concepción (utilizando el framework para identificar problemas Big Data), hasta la construcción de un prototipo que implemente la arquitectura de referencia descrita previamente.

IX. CONCLUSIONES

En este artículo se realizó un estado del arte general sobre el análisis de Big Data partiendo de los posibles hechos que motivan a los analistas a clasificar un problema dentro de esta área de estudio, y teniendo en cuenta la importante evolución que han presentado los métodos tradicionales de Data Warehousing, el proceso del KDD y los gestores de datos. En este artículo se mostró el enfoque principal que diferentes

autores han dado a Big Data (Las 3 Vs) y se incluyó una cuarta V para profundizar sobre el proceso de Inteligencia de Negocios sobre Big Data (Big Data Analytics), describiendo las aplicaciones, oportunidades, retos y tendencias que se presentan actualmente en la materia, basados en un amplio seguimiento bibliográfico.

Big Data no hace referencia únicamente a los datos, sino que también comprende todo el espectro de técnicas, métodos, herramientas y tecnologías alternativas, que permiten resolver problemas que involucran cierta complejidad, de una forma más eficaz y eficiente que los métodos tradicionales. A pesar de que con Big Data se puede lograr estos beneficios (eficacia y eficiencia) en problemas complejos de análisis de datos, Big Data Analytics no está desplazando el Data Warehousing tradicional. En lugar de esto, ambos métodos (los tradicionales y los relacionados con Big Data) se están complementando en pro de implementar soluciones que involucren todos los escenarios posibles.

Con respecto a las oportunidades de Big Data Analytics, se puede concluir que la revolución de los datos descrita en el artículo ha generado mayores ventajas y beneficios en diversos sectores, como la salud, la ciencia, los negocios y el gobierno, los cuales han permitido mejorar la calidad de vida de las personas y contribuir con el desarrollo de las regiones que hacen uso de la misma. Sin embargo, esta revolución de datos también ha traído nuevos desafíos que no se contemplaban en los métodos tradicionales, que van desde la captura y almacenamiento de los datos, hasta su análisis e interpretación. Adicionalmente, la cultura en las organizaciones es otro reto para Big Data Analytics, ya que estas deben ser conscientes de sus necesidades (estratégicas, económicas, funcionales) antes de abordar problemas de este tipo; de lo contrario, dichas implementaciones podrían fracasar.

Para abordar los desafíos que involucra una implementación de Big Data Analytics, diferentes autores, comunidades y organizaciones en general, han estudiado y propuesto tendencias en técnicas, métodos, herramientas y tecnologías, que permiten realizar implementaciones más transparentes. Sobre estas tendencias, en este artículo se puede concluir que, a pesar de que Big Data no es Hadoop (y visceversa), este framework se encuentra incluido en gran parte de las implementaciones de Big Data por el óptimo manejo que éste da a la distribución, tanto de datos como tareas. También se puede concluir que Cloud Computing y NoSQL aparecen constantemente dentro de las implementaciones de Big Data como apalancadores de la idea de obtener valor de los datos a bajo costo y en tiempos de respuesta igualmente bajos. En estos dos últimos puntos, cabe destacar que existe una gran tendencia relacionada con el surgimiento de organizaciones que ofrecen Big Data Analytics como servicio (o DaaS). Sin embargo, los líderes mundiales en soluciones analíticas como Oracle, Teradata, SAP, IBM y Microsoft no han sido desplazados ya que éstos se encuentran en constante evolución a la par de estas nuevas tecnologías.

RECONOCIMIENTO

Agradecimientos a los Ingenieros Iván Mauricio Cabezas Troyano, PhD y Luis Merchán Paredes, PhD, quienes motivaron este estudio por medio de sus cátedras impartidas dentro del contexto académico y sus experiencias personales y profesionales. Con sus aportes técnicos y metodológicos hicieron posible que la investigación estuviese centrada en proporcionar un marco bajo el cual se puedan adelantar estudios posteriores.

REFERENCIAS

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Networks*, vol. 54, pp. 2787–2805, 2010.

[2] K. Krishnan, *Data Warehousing in the Age of Big Data*. 2013, p. 371.

[3] C. L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny).*, pp. 1–34, Jan. 2014.

[4] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable : A Distributed Storage System for Structured Data," *ACM Trans. Comput. Syst.*, vol. 26, pp. 1–26, 2008.

[5] G. Decandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo : Amazon ' s Highly Available Key-value Store," *October*, vol. 41, pp. 205–220, 2007.

[6] J. Dean and S. Ghemawat, "MapReduce : Simplified Data Processing on Large Clusters," *Commun. ACM*, vol. 51, no. 1, pp. 1–13, 2008.

[7] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google file system," *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5. p. 29, 2003.

[8] "Apache™ Hadoop®!" [Online]. Available: http://hadoop.apache.org/. [Accessed: 18-May-2014].

[9] E. Letouzé, "Big Data for Development : Challenges & Opportunities," no. May 2012, 2012.

[10] "Gartner: Top 10 Strategic Technology Trends For 2013." [Online]. Available: http://www.forbes.com/sites/ericsavitz/2012/10/23/gartner-top-10-strategic-technology-trends-for-2013/. [Accessed: 15-May-2014].

[11] "Global Leader in Data Warehousing, Big Data Analytic Technologies & Data Driven Marketing - Teradata." [Online]. Available: http://www.teradata.com/?LangType=1033. [Accessed: 14-Aug-2014].

[12] "allAfrica.com: Africa: Fight Poverty - With Data." [Online]. Available: http://allafrica.com/stories/201307101239.html?viewall=1. [Accessed: 11-May-2014].

[13] K. Normandeau, "Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity," *Inside Big Data*, 2013.

[14] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," *J. Sch. Psychol.*, vol. 19, pp. 51–56, 2005.

[15] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, “Trends in big data analytics,” *J. Parallel Distrib. Comput.*, Feb. 2014.

[16] S. Pandey and S. Nepal, “Cloud computing and scientific applications - Big data, scalable analytics, and beyond,” *Futur. Gener. Comput. Syst.*, vol. 29, no. 7, pp. 1774–1776, Sep. 2013.

[17] A. O’Driscoll, J. Daugelaite, and R. D. Sleator, “‘Big data’, Hadoop and cloud computing in genomics,” *J. Biomed. Inform.*, vol. 46, no. 5, pp. 774–781, Oct. 2013.

[18] D. Loshin, *Big Data Analytics From Strategic Planning to Enterprise Integration with Tools , Techniques , NoSQL and Graphs*. .

[19] C. Güemes, “Data Analytics as a Service: unleashing the power of Cloud and Big Data,” 2013.

[20] E. F. Codd, “A relational model of data for large shared data banks. 1970.,” *MD. Comput.*, vol. 15, no. 3, pp. 162–6, 1970.

[21] M. M. Astrahan, J. W. Mehl, G. R. Putzolu, I. L. Traiger, B. W. Wade, V. Watson, M. W. Blasgen, D. D. Chamberlin, K. P. Eswaran, J. N. Gray, P. P. Griffiths, W. F. King, R. A. Lorie, and P. R. McJones, “System R: relational approach to database management,” *ACM Transactions on Database Systems*, vol. 1. pp. 97–137, 1976.

[22] “Magic Quadrant for Data Warehouse Database Management Systems.” [Online]. Available: http://www.gartner.com/technology/reprints.do?id=1-1ROSS0X&ct=140310&st=sb. [Accessed: 06-May-2014].

[23] “Corporate Information Factory (CIF) Resources by Bill Inmon, Inmon Data Systems.” [Online]. Available: http://www.inmoncif.com/home/. [Accessed: 06-May-2014].

[24] Ralph Kimball, *The Data Warehouse Lifecycle Toolkit Table of Contents*, 1st ed. 1998, p. 771.

[25] “Full Circle: Decision Intelligence (DSS 2.0) by Claudia Imhoff, Colin White - BeyeNETWORK.” [Online]. Available: http://www.b-eye-network.com/view/8385. [Accessed: 07-May-2014].

[26] J. Celko, *JOE CELKO’S COMPLETE GUIDE TO NoSQL: What every SQL professional needs to know about nonrelational databases*, 1st ed. Waltham, USA: Elsevier Inc., 2014, p. 227.

[27] A. Schram and K. M. Anderson, “MySQL to NoSQL Data Modeling Challenges in Supporting Scalability,” pp. 191–202.

[28] E. Barbierato, M. Gribaudo, and M. Iacono, “Performance evaluation of NoSQL big-data applications using multi-formalism models,” *Futur. Gener. Comput. Syst.*, Jan. 2014.

[29] J. Pokorny, “NoSQL databases: a step to database scalability in web environment,” *Int. J. Web Inf. Syst.*, vol. 9, no. 1, pp. 69–82, 2013.

[30] S. K. Gajendran, “A Survey on NoSQL Databases,” 1998.

[31] “Voldemort.” [Online]. Available: http://www.project-voldemort.com/voldemort/. [Accessed: 11-May-2014].

[32] “memcached - a distributed memory object caching system.” [Online]. Available: http://memcached.org/. [Accessed: 11-May-2014].

[33] “Redis.” [Online]. Available: http://redis.io/. [Accessed: 11-May-2014].

[34] “Riak | Basho Technologies.” [Online]. Available: http://basho.com/riak/. [Accessed: 11-May-2014].

[35] “MongoDB.” [Online]. Available: http://www.mongodb.org/. [Accessed: 11-May-2014].

[36] “Apache CouchDB.” [Online]. Available: http://couchdb.apache.org/. [Accessed: 11-May-2014].

[37] “The Apache Cassandra Project.” [Online]. Available: http://cassandra.apache.org/. [Accessed: 11-May-2014].

[38] “HBase - Apache HBase™ Home.” [Online]. Available: https://hbase.apache.org/. [Accessed: 11-May-2014].

[39] “Neo4j - The World’s Leading Graph Database.” [Online]. Available: http://www.neo4j.org/. [Accessed: 11-May-2014].

[40] “Gartner IT Glossary.” [Online]. Available: http://www.gartner.com/it-glossary. [Accessed: 07-May-2014].

[41] “Amazon Web Services (AWS) – Servicios de informática en la nube.” [Online]. Available: http://aws.amazon.com/es/. [Accessed: 19-May-2014].

[42] “Gartner: 10 Critical Tech Trends For The Next Five Years.” [Online]. Available: http://www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years/. [Accessed: 15-May-2014].

[43] Z. Qiu, Z. W. Lin, and Y. Ma, “Research of Hadoop-based data flow management system,” *J. China Univ. Posts Telecommun.*, vol. 18, no. SUPPL.2, pp. 164–168, Dec. 2011.

[44] “Apache Ambari.” [Online]. Available: http://ambari.apache.org/. [Accessed: 27-Aug-2014].

[45] “Apache Hive TM.” [Online]. Available: http://hive.apache.org/. [Accessed: 20-May-2014].

[46] “Oozie - Apache Oozie Workflow Scheduler for Hadoop.” [Online]. Available: http://oozie.apache.org/. [Accessed: 27-Aug-2014].

[47] “Apache Sqoop.” [Online]. Available: http://sqoop.apache.org/. [Accessed: 27-Aug-2014].

[48] “Welcome to Apache Pig!” [Online]. Available: http://pig.apache.org/. [Accessed: 20-May-2014].

[49] “Apache ZooKeeper - Home.” [Online]. Available: http://zookeeper.apache.org/. [Accessed: 27-Aug-2014].

[50] “Welcome to Apache Flume — Apache Flume.” [Online]. Available: http://flume.apache.org/. [Accessed: 22-Aug-2014].

[51] "Welcome to Apache Avro!" [Online]. Available: http://avro.apache.org/. [Accessed: 27-Aug-2014].

[52] "Apache Mahout: Scalable machine learning and data mining." [Online]. Available: https://mahout.apache.org/. [Accessed: 22-Aug-2014].

[53] "Spark Streaming | Apache Spark." [Online]. Available: https://spark.apache.org/streaming/. [Accessed: 22-Aug-2014].

[54] "Pentaho | Business analytics and business intelligence leaders." [Online]. Available: http://www.pentaho.com/. [Accessed: 07-Sep-2014].

[55] "SAP Software & Solutions | Business Applications & IT | SAP." [Online]. Available: http://www.sap.com/index.html. [Accessed: 16-Jul-2014].

[56] "Microsoft US | Devices and Services." [Online]. Available: http://www.microsoft.com/en-us/default.aspx. [Accessed: 07-Sep-2014].

[57] "Jaspersoft Business Intelligence." [Online]. Available: https://www.jaspersoft.com/es. [Accessed: 07-Sep-2014].

[58] "IBM - Cognos Business Intelligence ." IBM Corporation, 09-Sep-2014.

[59] "Oracle Business Intelligence Enterprise Edition | Oracle Technology Network | Oracle." [Online]. Available: http://www.oracle.com/technetwork/middleware/bi-enterprise-edition/overview/index.html. [Accessed: 10-Sep-2014].

[60] "High Level Panel - The Post 2015 Development Agenda." [Online]. Available: http://www.post2015hlp.org/. [Accessed: 11-May-2014].

[61] A. H. B. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute*, 2011. [Online]. Available: http://scholar.google.com/scholar.bib?q=info:kkCtazs1Q6wJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,47&ct=citation&cd=0.

[62] "FACT SHEET: Big Data and Privacy Working Group Review | The White House." [Online]. Available: http://www.whitehouse.gov/the-press-office/2014/05/01/fact-sheet-big-data-and-privacy-working-group-review. [Accessed: 28-Aug-2014].

[63] "Evolución de la gripe en Google." [Online]. Available: http://www.google.org/flutrends/. [Accessed: 28-Aug-2014].

[64] "Google Trends del dengue." [Online]. Available: https://www.google.org/denguetrends/intl/es_419/. [Accessed: 28-Aug-2014].

[65] "Ciudad creativa digital." [Online]. Available: http://www.carloratti.it/FTP/CCD/. [Accessed: 28-Aug-2014].

[66] "Ubidots - Internet of Things application development platform." [Online]. Available: http://ubidots.com/. [Accessed: 28-Aug-2014].

[67] "CoCoRaHS - Community Collaborative Rain, Hail & Snow Network." [Online]. Available: http://www.cocorahs.org/. [Accessed: 08-Sep-2014].

[68] "IPython - Interactive Computing." [Online]. Available: http://ipython.org/. [Accessed: 08-Sep-2014].

[69] "UIT: Comprometida para conectar el mundo." [Online]. Available: http://www.itu.int/es/Pages/default.aspx. [Accessed: 07-Sep-2014].

[70] "Microsoft US | Devices and Services." [Online]. Available: http://www.microsoft.com/en-us/default.aspx. [Accessed: 07-Sep-2014].

[71] "Jaspersoft Business Intelligence." [Online]. Available: https://www.jaspersoft.com/es. [Accessed: 07-Sep-2014].

[72] "Pentaho | Business analytics and business intelligence leaders." [Online]. Available: http://www.pentaho.com/. [Accessed: 07-Sep-2014].

[73] "Business Intelligence and Analytics | Tableau Software." [Online]. Available: http://www.tableausoftware.com/. [Accessed: 08-Sep-2014].

[74] "Top 10 Data Warehousing Trends and Opportunities for 2014," 2014.

[75] "Oracle | Hardware and Software, Engineered to Work Together." [Online]. Available: http://www.oracle.com/index.html. [Accessed: 16-Jul-2014].

[76] "SAP Software & Solutions | Business Applications & IT | SAP." [Online]. Available: http://www.sap.com/index.html. [Accessed: 16-Jul-2014].

[77] "Azure: Plataforma en la nube de Microsoft | Hospedaje en la nube | Servicios en la nube." [Online]. Available: http://azure.microsoft.com/es-es/. [Accessed: 16-Jul-2014].

[78] "IBM - United States." IBM Corporation, 13-May-2014.

[79] "Big Data Discovery and Data Sharing | 1010data." [Online]. Available: http://www.1010data.com/. [Accessed: 11-May-2014].

[80] "AWS | Amazon DynamoDB – NoSQL Database Service." [Online]. Available: https://aws.amazon.com/es/dynamodb/. [Accessed: 11-May-2014].

[81] "The Platform for Big Data and the Leading Solution for Apache Hadoop in the Enterprise - Cloudera." [Online]. Available: http://www.cloudera.com/content/cloudera/en/home.html. [Accessed: 18-May-2014].

[82] "Enterprise NoSQL Database | MarkLogic." [Online]. Available: http://www.marklogic.com/. [Accessed: 19-May-2014].

[83] "Supercomputing for Data Science - Big Data Analytics - Kognitio." [Online]. Available: http://www.kognitio.com/. [Accessed: 19-May-2014].

[84] "Big Data Analytics | Transforming Data Into Value | Actian." [Online]. Available: http://www.actian.com/. [Accessed: 11-May-2014].

[85] "Home | Pivotal." [Online]. Available: http://www.gopivotal.com/. [Accessed: 16-Jul-2014].

[86] E. Begoli, "A Short Survey on the State of the Art in Architectures and Platforms for Large Scale Data Analysis and Knowledge Discovery from Data."

[87] A. Abouzeid and K. Bajda-pawlikowski, "HadoopDB : An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads."

[88] "PostgreSQL: The world's most advanced open source database." [Online]. Available: http://www.postgresql.org/. [Accessed: 21-Aug-2014].

[89] "Real-Time Analytics Platform | Big Data Analytics | MPP Data Warehouse." [Online]. Available: http://www.vertica.com/. [Accessed: 21-Aug-2014].

[90] "IBM - Software - Information Management - Netezza - Colombia." IBM Corporation, 06-Feb-2014.

[91] "Big Data | Pivotal Greenplum Database | Pivotal." [Online]. Available: http://www.pivotal.io/big-data/pivotal-greenplum-database. [Accessed: 22-Aug-2014].

[92] "Enterprise Data Management, Analysis and Mobilization Software - Sybase Inc." [Online]. Available: http://www.sybase.com/. [Accessed: 22-Aug-2014].

[93] "Business Analytics and Business Intelligence Software | SAS." [Online]. Available: http://www.sas.com/en_us/home.html. [Accessed: 22-Aug-2014].

[94] "EXASOL | www.exasol.com." [Online]. Available: http://www.exasol.com/en/. [Accessed: 18-May-2014].

[95] "Apache Mahout: Scalable machine learning and data mining." [Online]. Available: https://mahout.apache.org/. [Accessed: 22-Aug-2014].

[96] "GraphLab.org | GraphLab Open Source." [Online]. Available: http://graphlab.org/projects/index.html. [Accessed: 22-Aug-2014].

[97] M. Hall, H. National, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software : An Update," *SIGKDD Explor.*, vol. 11, pp. 10–18, 2009.

[98] "Google BigQuery - Fully Managed Big Data Analytics Service — Google Cloud Platform — Google Cloud Platform." [Online]. Available: https://cloud.google.com/products/bigquery/. [Accessed: 22-Aug-2014].

[99] "Apache Drill." [Online]. Available: http://incubator.apache.org/drill/. [Accessed: 22-Aug-2014].

[100] S. Melnik, A. Gubarev, J. J. Long, G. Romer, S. Shivakumar, M. Tolton, and T. Vassilakis, "Dremel : Interactive Analysis of Web-Scale Datasets."

[101] K. Goodhope, J. Koshy, and J. Kreps, "Building LinkedIn's Real-time Activity Data Pipeline.," *IEEE Data Eng*, pp. 1–13, 2012.

[102] L. Neumeyer and B. Robbins, "S4 : Distributed Stream Computing Platform."

[103] "Storm, distributed and fault-tolerant realtime computation." [Online]. Available: https://storm.incubator.apache.org/. [Accessed: 22-Aug-2014].

[104] "Microsoft SQL Server." [Online]. Available: http://www.microsoft.com/OEM/es/products/servers/Pages/sql_server.aspx#fbid=s-UKc1guews. [Accessed: 24-Aug-2014].

[105] "Tim Berners-Lee. Linked Data - Design Issues." [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html. [Accessed: 24-Aug-2014].

[106] "Conocimiento – Dentro de Google – Google." [Online]. Available: http://www.google.com/insidesearch/features/search/knowledge.html. [Accessed: 24-Aug-2014].

[107] "DBPedia Knowledge Base." [Online]. Available: http://dbpedia.org/About. [Accessed: 24-Aug-2014].

[108] "Freebase." [Online]. Available: https://www.freebase.com/. [Accessed: 24-Aug-2014].

[109] "Classora Knowledge Base." [Online]. Available: http://www.classora.com/. [Accessed: 24-Aug-2014].

[110] "Google App Engine - Platform As A Service & Application Hosting — Google Cloud Platform — Google Cloud Platform." [Online]. Available: https://cloud.google.com/products/app-engine/?utm_source=google&utm_medium=cpc&utm_campaign=appengine-search-global&gclid=CLXE7Iqmq8ACFUVo7AodqSkAkg. [Accessed: 24-Aug-2014].