DISEÑO DE MODELO TECNOLÓGICO PARA EL USO DE BIG DATA EN EL ANÁLISIS Y VISUALIZACIÓN DE INFORMACIÓN PARA LA PEQUEÑA Y MEDIANA EMPRESA (PYMES)

Yancy Steffany Ventura Aguilar

José Guillermo Rivera Pleitez

Saúl Antonio Cornejo Hernández

Facultad de Ingeniería

CONTENIDO

Introducción

Big Data ha sido muy usado en el medio de la informática y de las grandes empresas, ya que en ellas se puede visualizar la gran cantidad de información que se maneja hoy en día, es tanta la información que entra y sale que es un reto el manejo de esa información. (Rosa y Rivera Pleitez, 2016)

Big Data es un término que hace referencia a una cantidad de datos tal que supera la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable. El volumen de los datos masivos crece constantemente. En 2012 se estimaba su tamaño de entre una docena de terabytes hasta varios petabytes de datos en un único conjunto de datos. Se continúan usando datos masivos y en mayor escala que hace 14 años, por lo tanto, para las empresas se hace necesario buscar herramientas que permitan dar soluciones a la demanda de grandes cantidades de datos, para el procesamiento y análisis como el caso de MapR, Cyttek Group, Cloudera, Hadoop, entre otros. (Rosa y Rivera Pleitez, 2016)

Por lo tanto, la realidad no se puede cambiar si no se debe orientar en la misma dirección los avances de la ciencia y tecnología, por lo que existe la necesidad de trabajar con una gran cantidad de datos; pero la gran mayoría de empresas no saben cómo hacerlo.

Esta investigación servirá de referencia para dar a conocer el uso de herramientas de Big Data en El Salvador. Es un país muy pequeño en extensión territorial y población en comparación con el resto de países del mundo.

En cuanto a la tecnología, se trata de ir a la vanguardia; sobre todo, en temas como las telecomunicaciones, pero el concepto de Big Data es algo novedoso, aunque con muchas ganas de incursionar en el uso de las herramientas que esto conlleva, pues las empresas se preguntan cómo procesar y almacenar grandes volúmenes de datos y luego analizarlos. (Rosa y Rivera Pleitez, 2016)

Es tanta la información que se genera a diario en la web a través de las redes sociales, buscadores, almacenamiento de datos en la nube, entre otras. Por lo que resulta abrumador y solo el hecho de saber cómo se consigue captar y analizar dicha información es sorprendente. (Rosa y Rivera Pleitez, 2016)

También se sabe que las redes sociales hoy en día, aportan mucha información relevante que los usuarios comparten libremente y públicamente en la web. Para los que están inmersos en este medio, no es desconocido que a muchas personas les encanta publicar los lugares en los que están en un momento dado, las marcas que prefieren, ropa, zapatos, accesorios, perfumes, comidas, restaurantes, entre otras. (Rosa y Rivera Pleitez, 2016)

Todo esto es aprovechable por las empresas, para detectar tendencias en el mercado y enfocar las acciones que se van a llevar a cabo, algo que ayuda a tomar mejores decisiones y a que los resultados sean mejores. (Rosa y Rivera Pleitez, 2016)

Por supuesto, las ventajas las obtendrán aquellas empresas que sepan cómo procesar y analizar esos datos y es allí donde muchas se quedan estancadas en seguir haciendo los procedimientos cotidianos, por la ignorancia del uso de herramientas que facilitarían el procesado masivo de datos y en poco tiempo. (Rosa y Rivera Pleitez, 2016)

Por otro lado, están los dataset públicos –distintos formatos– y es allí donde surge el problema, cuando los datos no son estructurados como comúnmente se acostumbra utilizarlos en las bases de datos relacionales tradicionales, pues se encuentran en formatos como json, csv, dat, arff, ncol, etc. En estos casos, se hace necesario el uso de herramientas que permitan almacenar y procesar ese tipo de ficheros. (Rosa y Rivera Pleitez, 2016)

Por esto se consideró importante una propuesta metodológica que haga uso de herramientas propias de Big Data para el procesamiento masivo de información, análisis de los resultados y visualización de los datos.

Lo que se pretendió con esa información, es ayudar a la pequeña empresa a conocer los procedimientos necesarios y las herramientas que serán útiles para solventar el problema de trabajar con datos masivos y obtener resultados en menor tiempo.

Se tomó a bien hacer uso de dos dataset para aprovechar el uso de las herramientas y mostrar lo que se puede hacer con la información que contienen. En este trabajo se explica la identificación del problema, planteando la necesidad actual, trabajando con una prueba diagnóstica aplicada a los participantes del estudio.

A partir de la necesidad identificada se realizó la propuesta de innovación con la que se esperaba dar a conocer las herramientas que ayudarían al manejo de grandes volúmenes de información para la pequeña y mediana empresa.

Con una propuesta adecuada a los recursos de la pequeña y mediana empresa los resultados esperados fueron aplicables a los procesos metodológicos y técnicos para facilitaran la aplicación de Big Data por medio de herramientas efectivas generando información para la toma de decisiones.

CAPÍTULO I. LA NECESIDAD DE INNOVAR

Necesidad y problemas asociados

El comercio electrónico actualmente es una realidad que permite a las empresas ofrecer sus productos y servicios a una mayor cantidad de potenciales consumidores; lo que permite aumentar sus ingresos al incrementar sus transacciones, respecto aquellas empresas que no utilizan este medio para hacer negocios. (Centro de Comercio Internacional UNCTAD/OMC Colombia. 2002)

Es necesario indicar que las condiciones están dadas actualmente en el país para que las empresas utilicen este medio para ofrecer sus productos, servicios y realizar comercio entre los actores involucrados. Estos elementos son: seguridad de la información, tecnología, medios transaccionales para operatividad, estrategias de marketing, entre otros. Sin embargo, existe un desconocimiento que genera incertidumbre para utilizar este medio para hacer negocios. (Centro de Comercio Internacional UNCTAD/OMC Colombia. 2002)

Algunas empresas poseen algunos de los indicadores mencionados anteriormente para realizar el comercio, pero se resiente que no se logra cumplir satisfactoriamente las bondades que el comercio electrónico genera. (Rosa y Rivera Pleitez, 2016)

La pequeña y mediana empresa salvadoreña– conocida como pymes– están interesadas en conocer cuáles son las necesidades reales de sus potenciales clientes, con el objeto de llegar a ofrecerle productos y servicios más específicos y de ser posibles personalizados. Esta acción es muy compleja realizarla con un comercio tradicional, es complicada a través de comercio electrónico tradicional; pero es posible a través de comercio electrónico, con la conjunción de todos los indicadores mencionados anteriormente, alineados para el análisis de la información que proporcionan los potenciales clientes. (Rosa y Rivera Pleitez, 2016)

Si la información que se genera en la web es analizada por medio de herramientas como Big Data; el resultado de ese análisis genera oportunidades al comercio electrónico, y es donde las pymes puedan aprovechar estas oportunidades.

Las empresas siempre buscan obtener información valiosa para tomar las mejores decisiones en su negocio, para ello existen muchas herramientas y tecnologías que les aportan grandes beneficios. Se trata de la tecnología Big Data que fomenta la eficiencia, calidad y los productos y servicios personalizados, lo que produce niveles altos de experiencia y satisfacción hacia sus clientes. (Rosa y Rivera Pleitez, 2016)

Para este proyecto se realizó un diagnostico con el objetivo de conocer sobre el uso del Big Data, las oportunidades de negocio en las pymes. El estudio fue de tipo descriptivo por medio de una encuesta que abarca únicamente a las empresas dedicadas al rubro de servicios de San Salvador para la pequeña y mediana empresa, cuya muestra se seleccionó de manera aleatoria para dar oportunidad a la mayor cantidad de empresas.

Participantes

El proceso de recolección de datos se realizó en las pequeñas y medianas empresas pymes dedicadas a rubro de servicios en el departamento de San Salvador. En los anexos se describe el listado de empresas encuestadas. La encuesta se distribuyó a 90 empresas de las cuales contestaron 72.

Instrumento

En la encuesta se buscó conocer la infraestructura tecnológica con las que cuentan actualmente las empresas y se dará a conocer si las empresas hacen uso de la tecnología de Big Data para el manejo de toma de decisiones.

La encuesta se pasó a las empresas en enero de 2016, dicho instrumento utilizado para la recolección de información constó de 17 preguntas que se presentan en el diseño de la encuesta.

Procedimientos

Lo primero que se realizó fue enviar con herramientas en línea la encuesta a la empresa, este proceso se efectuó durante la primera semana de enero 2016.

Luego de pasar las encuestas se tabularon los resultados entre la segunda y tercer semana de enero de 2016, para obtener los resultados que permiten realizar un análisis con el que se establecieron conclusiones que describió información importante para medir el grado de utilización y aprovechamiento de la tecnología Big Data por parte de las pymes.

Se enviaron un total de 90 encuestas a los representantes de mediana y pequeña empresa para obtener el mayor número de resultados, de estas un 80 % respondieron en un periodo de tres semanas. Lo que permitió el análisis del diagnóstico inicial. El 80 % equivale a 72 empresas que contestaron la encuesta en línea.

Resultados del Diagnóstico

Con la encuesta realizada se buscó medir el grado de utilización y aprovechamiento de la tecnología Big Data en el comercio electrónico realizado por las pymes. A continuación se presentan los resultados obtenidos de las encuestas aplicadas durante el mes de enero de 2016.

En la Figura 1. se observa el análisis porcentual del género de los entrevistados en el estudio diagnóstico:

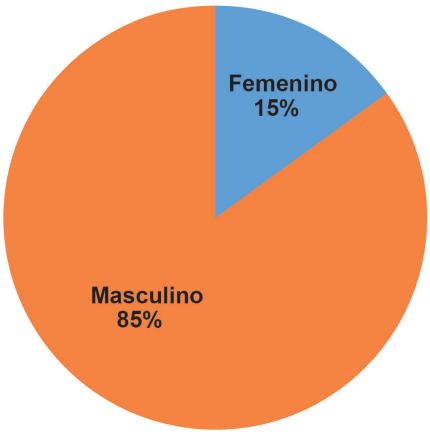


Figura 1. Género de los encuestados.
Fuente: Propia.

Los resultados en cuanto a las edades de los entrevistados (Figura 2) el mayor porcentaje osciló entre los 30 a más años de edad.

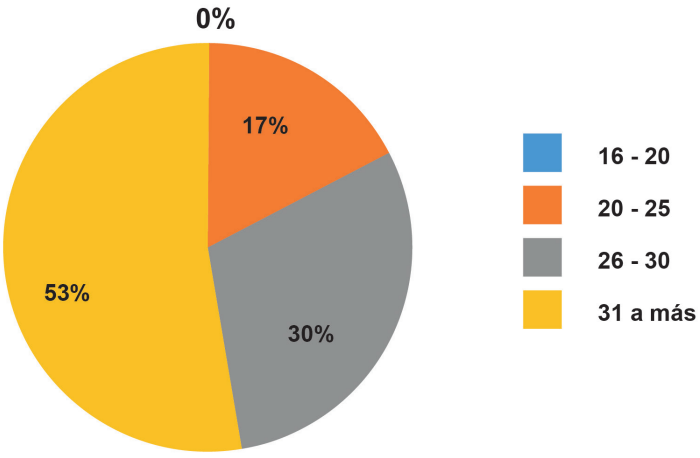


Figura 2. Edad de los encuestados.
Fuente: Propia.

En la Figura 3 se presenta la cantidad de empleados con los que cuentan las empresas encuestadas.

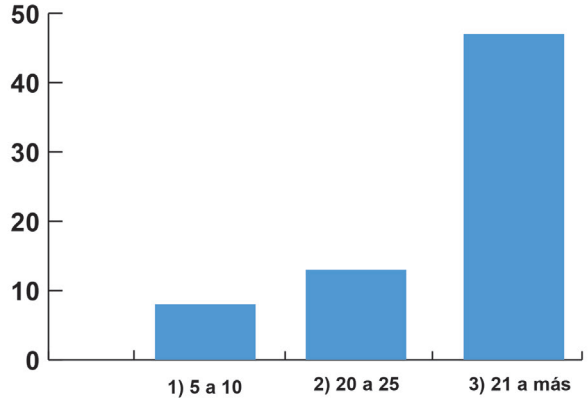


Figura 3. Cantidad de empleados de la empresa.
Fuente: Propia.

Los resultados obtenidos en la investigación con respecto a las empresas encuestadas que cuentan con algún tipo de tecnología de análisis arrojó que 42 empresas contaban con tecnología (Figura 4).

Muestra que la mayoría de las empresas dispone de tecnología, pero no está haciendo utilizada para el análisis de información, este es un indicador de la carencia de conocimiento o interés que estas empresas poseen en relación al uso de Big Data.

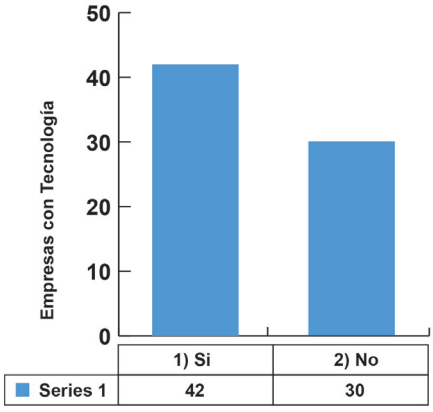


Figura 4. Presencia de tecnología en la empresa
Fuente: Propia.

Resultados obtenidos de la infraestructura tecnológica con la que cuentan las empresas entrevistadas:

Como observamos en el gráfico (Figura 5), el porcentaje mayor nos indica que las empresas encuestadas cuentan con al menos un tipo de infraestructura dentro de su entidad; sin embargo, el porcentaje es variante entre los diferentes tipos de infraestructura tecnológica.

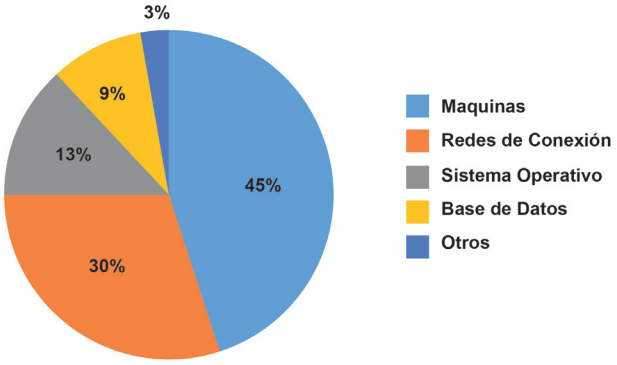


Figura 5. Infraestructura tecnológica que posee la empresa
Fuente: Propia.

Con respecto a los medios digitales de comunicación con los que cuentan las empresas encuestadas tenemos la siguiente gráfica de análisis (Figura 6).

Empresas que cuentan con las redes sociales están con un 35 % de porcentaje.

El segundo mayor porcentaje de la gráfica es resultado de las empresas que poseen su canal en comunicación interactiva, seguido por las empresas que poseen medios electrónicos.

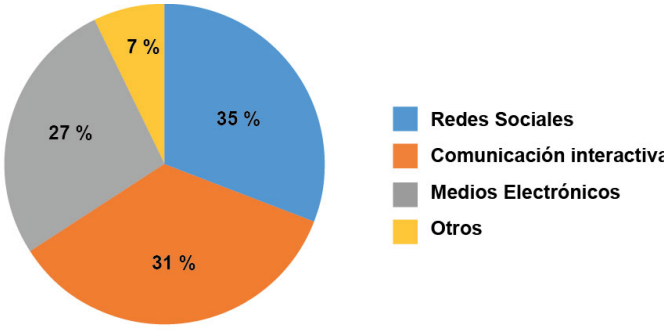


Figura 6. Medios digitales de comunicación en las empresas.
Fuente: Propia.

A continuación se muestran los resultados obtenidos de la frecuencia con la que las empresas encuestadas usan los medios de comercio electrónico, se representa en la siguiente gráfica (Figura 7).

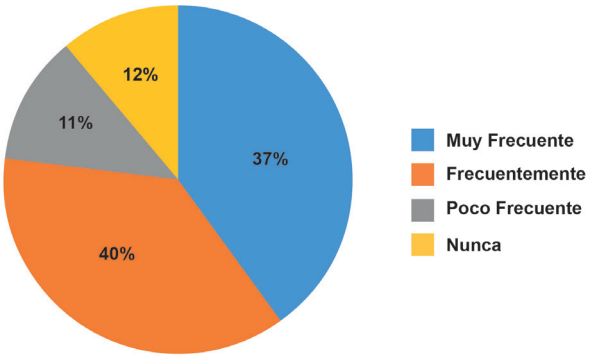


Figura 7. Frecuencia de uso de herramientas de internet para promocionar los productos de la empresa.
Fuente: Propia.

Analizando de forma porcentual las respuestas obtenidas de las empresas encuestadas detallamos en la siguiente gráfica (Figura 8).

El grado de aceptación en el uso de las redes sociales y otros para el crecimiento en las empresas encuestadas con un porcentaje de 89 %. El porcentaje de quienes, a su criterio, no consideran que los medios electrónicos de comunicación sean de ayuda para el crecimiento de su empresa es de un 89 % seguido de los que dudan con un 11 %.

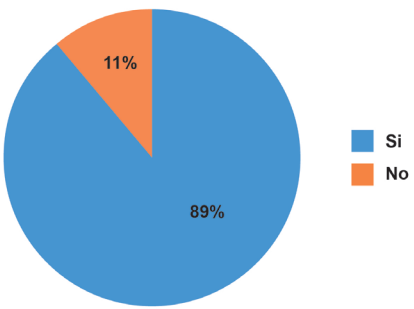


Figura 8. Importancia de las herramientas de internet para el aumento de clientes en la empresa.
Fuente: Propia.

Datos porcentuales obtenidos del estudio en las pymes base para otorgar descuentos en los productos y promocionar dentro de su empresa se muestra en la siguiente gráfica (Figura 9).

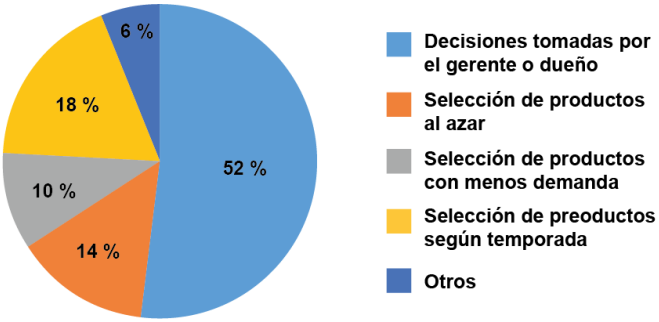


Figura 9. Forma de otorgar los descuentos y promoción de los productos en la empresa.

Fuente: Propia.

Resultados alcanzados para conocer las herramientas o técnicas utilizadas por las empresas encuestadas para dar a conocer productos nuevos o promociones a sus clientes, se observa en la Figura 10.

En la actualidad las redes sociales son comúnmente utilizadas para promocionar y ofertar productos o servicios a nivel mundial, como se muestra en el gráfico logrando el primer lugar con un porcentaje de 52 %, seguido de medios electrónico con un 21 % sobre el resto de herramientas.

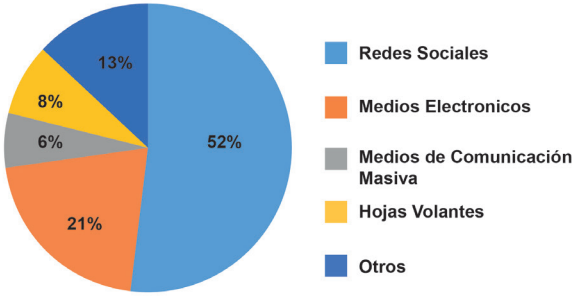


Figura 10. Herramientas de internet para dar a conocer los nuevos productos o promociones a los clientes existentes o potenciales.

Fuente: Propia.

En su mayoría, las empresas tomadas como muestras para este estudio, consideran que es de gran beneficio realizar análisis de datos para el desarrollo y crecimiento de su entidad. A continuación una gráfica que ejemplifica lo anterior con datos reales (Figura 11).

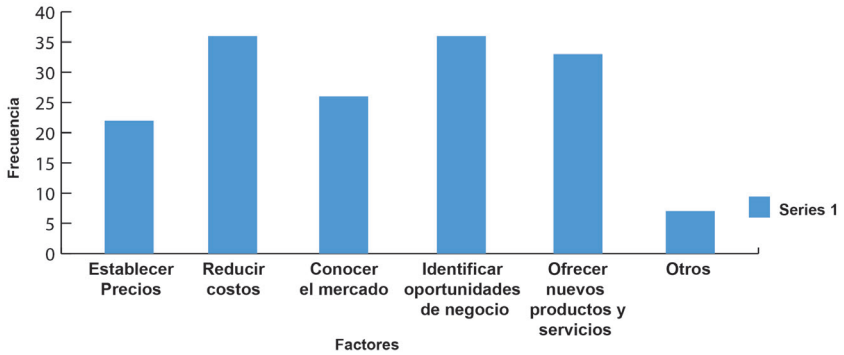


Figura 11. Factores que beneficiarían a la empresa si realizara análisis de datos.

Fuente: Propia.

De acuerdo a los resultados del estudio con respecto al análisis de datos, información de clientes, productos, promociones y datos importantes que lograrían ayudar a mejores resultados financieros a la empresa, se obtiene la siguiente gráfica (Figura 12).

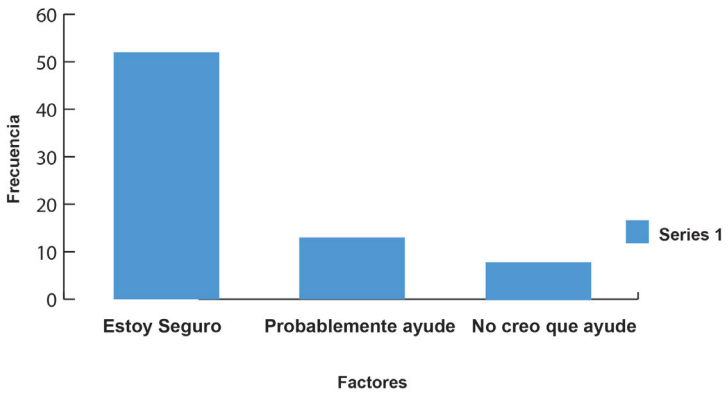


Figura 12. Consideración de cómo el análisis de datos, la información de los clientes, productos, promociones y datos importantes ayudan a la empresa.

Fuente: Propia.

Basado en el estudio del porcentaje de la toma de decisiones se obtiene el siguiente gráfico (Figura 13).

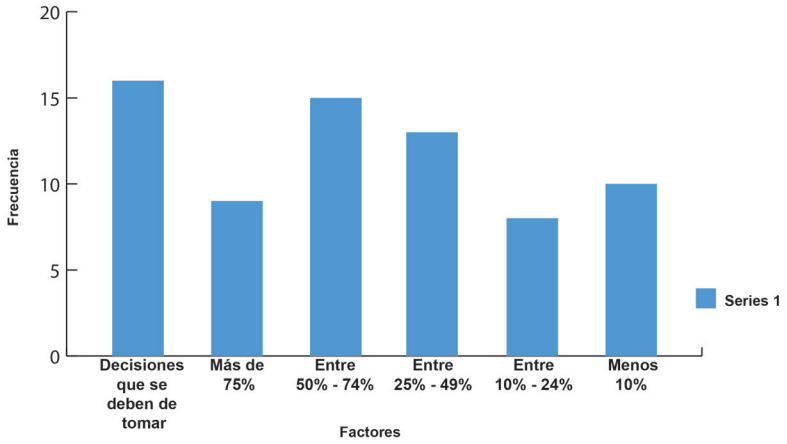


Figura 13. Decisiones tomadas con base en análisis de datos.
Fuente: Propia.

A continuación una representación gráfica en la Figura 14 muestra la ventaja competitiva dentro de las empresas encuestadas con referencia al análisis de dato. Como se observa en la siguiente gráfica, consideran ventaja competitiva el 49 %; seguido de relativamente buena con 34 %, los rangos más altos. Luego un 10 % considera que es muy poca venta, seguido de los que no lo consideran una ventaja con el 7 %.

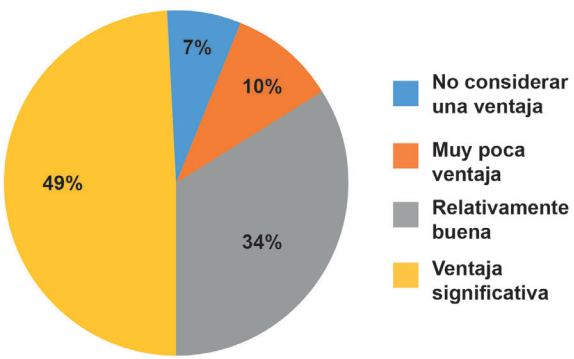


Figura 14. Medida como la empresa considera el análisis de datos como una ventaja competitiva.
Fuente: Propia.

Los resultados obtenidos de cuáles serían los factores que impedirían la implementación de un modelo de análisis de datos, según la investigación es la siguiente (Figura 15).

El mayor porcentaje contemplado 37 % considera que la falta de recursos financieros les impediría la implementación de una herramienta de análisis; seguido de un 20 %, que estiman que la falta de experiencia en el tema sería un impedimento.

En los siguientes rangos porcentuales se observa que la muestra tomada tiene el 18 % no sabe sobre el tema o no responde, seguido de un 14 % que considera no contar con la infraestructura necesaria, y para finalizar el 11 % considera que el análisis de dato e información no es aplicable para el rubro al que pertenece su empresa.

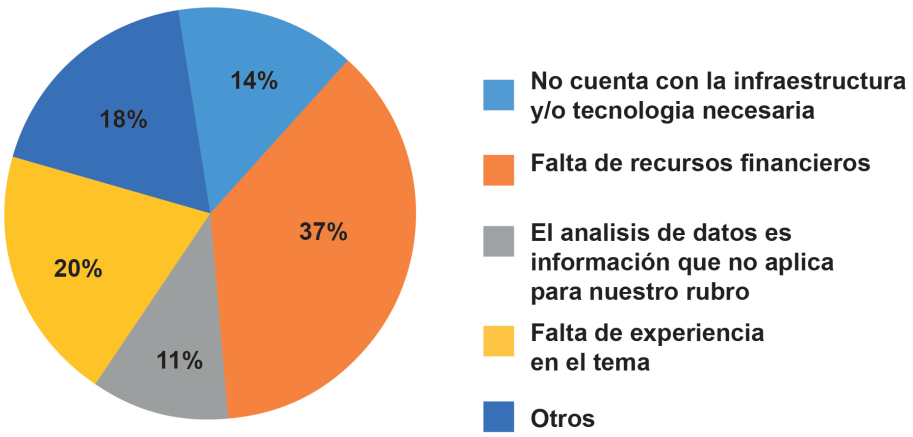


Figura 15. Factores que impedirían la implementación de un modelo de análisis de datos e información en la empresa.
Fuente: Propia.

El estudio muestra de cuántas empresas estarían interesadas en adquirir el software y herramienta de análisis dentro de la empresa en el siguiente gráfico (Figura 16).

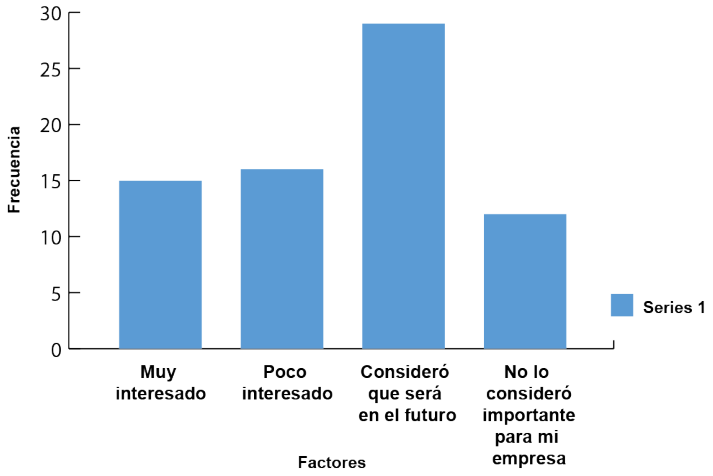


Figura 16. Interés en adquirir software tecnológico para realizar análisis de datos.
Fuente: Propia.

Los resultados obtenidos de los aspectos que consideran se deben tomar en cuenta para la adquisición de una herramienta para el análisis de datos de las empresas encuestadas, se representan en la siguiente gráfica (Figura 17).

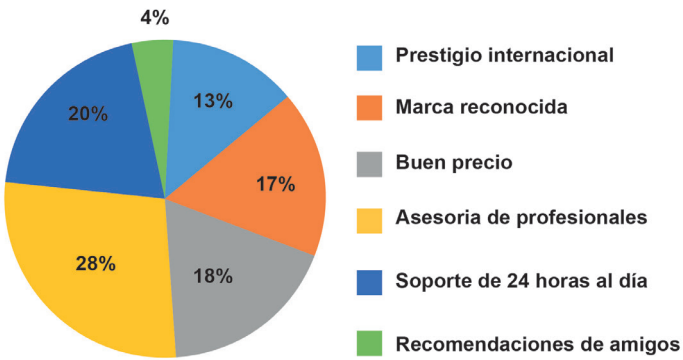


Figura 17. Aspectos a considerar para adquirir herramientas para el análisis de datos de la empresa.
Fuente: Propia.

Resultados obtenidos en la investigación con referencia a la cantidad en términos financieros que estarían dispuestos a pagar para adquirir un software de análisis de datos reflejado en la siguiente gráfica (Figura 18).

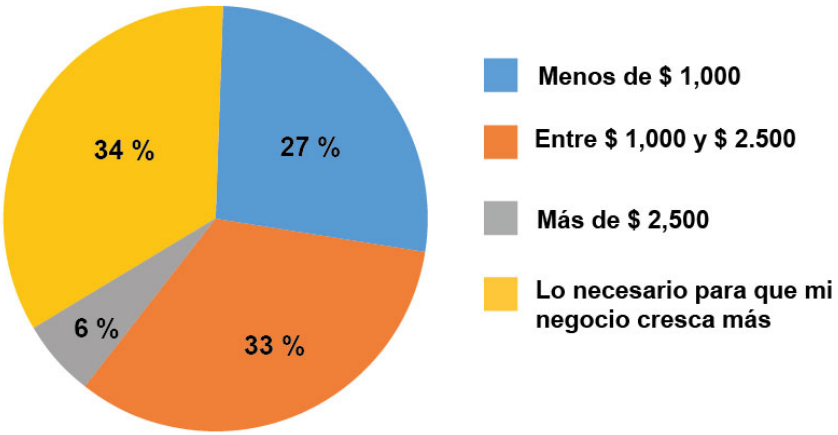


Figura 18. Inversión dispuesta a pagar para adquirir software de análisis de datos.
Fuente: Propia.

Áreas de la empresa que consideran deben ser capacitadas al adquirir un software para el análisis de datos se representa en la siguiente gráfica (Figura 19).

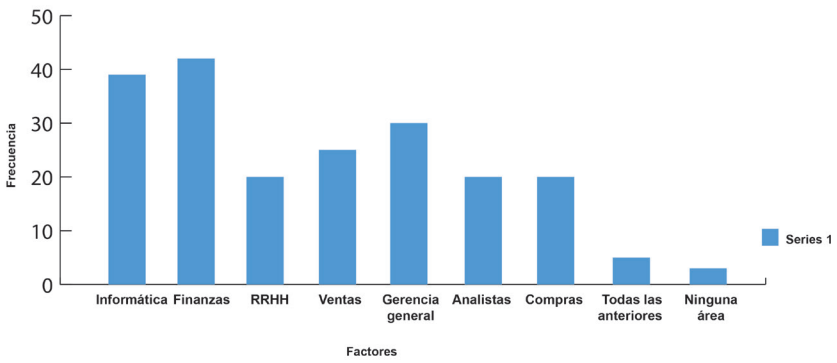


Figura 19. Capacitaciones necesarias para el análisis de datos por adquisición de software.
Fuente: Propia.

En el siguiente gráfico de muestra que los encuestados prefieren para la capacitación del personal en el uso de la herramienta del análisis de datos se realice dentro de su horario de trabajo. Muestra en el siguiente gráfico (Figura 20).

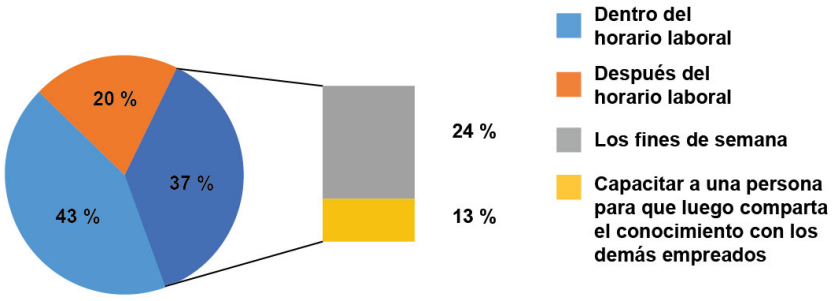


Figura 20. Horarios idóneos para capacitaciones en análisis de datos de la empresa.

Fuente: Propia.

Como se observa en la siguiente gráfica (Figura 21) la muestra entrevistada en solicitar firmar contrato de confidencialidad tiene un rango de 62, seguido de 51 que considera deben establecer políticas de seguridad. Seguido de los que consideran que el desarrollo de la herramienta es oportuno sea dentro de la empresa con un 9 y 8 los que solicitarían recomendaciones de otras empresas.

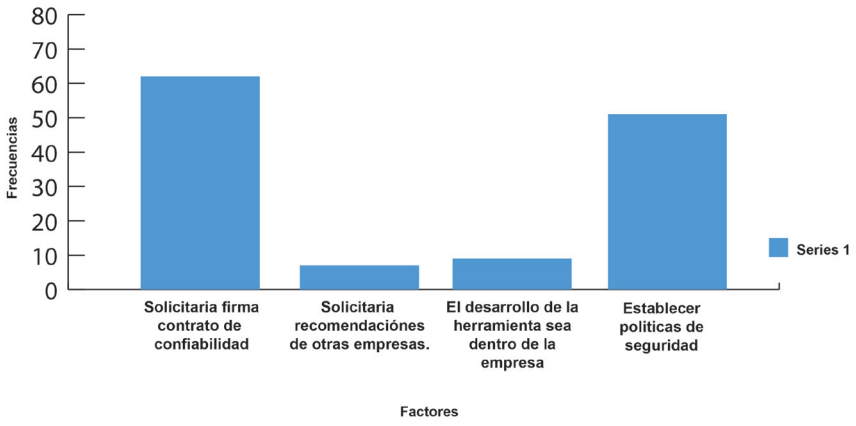


Figura 21. Medidas de seguridad para proteger la información de su negocio al contratar empresa creadora de software de análisis de datos.

Fuente: Propia.

Resultados de la prueba diagnóstica

Después de presentar los resultados se concluye lo siguiente: se han identificado los actores que se involucran en los procesos de comercio electrónico, con esto se facilitó la definición de los elementos que son parte del modelo de incorporación de la pymes al comercio electrónico.

Se determinó el escenario actual de las pymes, en el cual se puede hacer mención que el universo de información que se maneja en las instituciones es inmenso; los datos no están totalmente organizados lo que hace que las empresas no exploten óptimamente los segmentos de mercado que tienen y obtengan así los niveles de productividad óptimos.

Con los insumos de la prueba diagnóstica se demuestra la necesidad que existe de elaborar un diseño de procedimientos tecnológicos para el uso de Big Data para el almacenamiento, análisis y visualización de la información para la pymes.

B. Justificación

Muchas empresas, actualmente poseen infraestructura tecnológica, procesos sistematizados, estrategias de marketing; pero no son efectivas en capitalizar las bondades que ofrece la tecnología del internet, ya que no logran proporcionarle al potencial cliente, información oportuna y significativa sobre sus necesidades en el momento deseado, ya sea por una toma de decisiones tardía o por no disponer de la información necesaria para ofrecerle cuando se requiere. (Rosa y Rivera Pleitez, 2016)

Se pretendió que las pymes hagan uso de esta tecnología para que se incorporen al comercio electrónico, utilizando la información que se genera en la web, por lo que surge la necesidad de hablar de los fundamentos de Big Data, las herramientas para su uso, así como saber de dónde provienen los datos que proporcionan información sobre su negocio para permitir una mejor toma de decisiones.

El aprovechamiento de tecnología del Big Data permite que las pymes conozcan más de cerca a sus clientes, prestarle un mejor servicio, mejorar la calidad de sus productos, generar oportunidad para ingresar a nuevos mercados, completar su portafolio de clientes, entre otras tareas que generen beneficios al negocio a partir de información ya existente y que no es aprovechada.

Por lo tanto, la investigación sobre la tecnología de Big Data dentro de las pymes, estuvo basada en los siguientes indicadores: conocer la infraestructura tecnológica con las que se encuentran actualmente, la utilización y las herramientas que son parte esencial en el intercambio comercial de la misma, conocer los puntos en los

cuáles se puede mejorar, ampliar y/o remplazar aquellos que así lo requieran para poder agilizar los procesos de sus negocios.

Así mismo, la presente investigación pretende dar a conocer las herramientas necesarias para su posible implementación dentro de la empresa para el fortalecimiento tecnológico que permita una forma eficiente prestar servicios a sus clientes y que puedan verse reflejados en términos monetarios, entre otros.

C. Fundamentación teórica

A. Marco histórico

Las pymes en El Salvador

Descripción de las pymes

La pequeña y mediana empresa que normalmente se conoce como pymes es una empresa con características distintivas, y tiene dimensiones con ciertos límites ocupacionales y financieros prefijados por los estados o regiones. Las pymes son agentes con lógicas, culturas, intereses y un espíritu emprendedor específicos. A nivel nacional es uno de los actores claves del crecimiento salvadoreño, reconociéndole también como fuente directa de empleo, en el PIB y en el comercio exterior. La innovación tecnológica nacional e internacional en que se mueven los negocios, plantean importantes desafíos a las pymes. Se dividen en dos grandes segmentos: las sociedades y las empresas de hogares. (Martinez, 2013)

Las sociedades son entidades jurídicas en las cuales varias personas se asocian persiguiendo un fin común; entre estas se tienen sociedades anónimas, sociedades de responsabilidad limitada, sociedades en comandita, cooperativas de productores, entre otros. Las Empresas de hogares son empresas no constituidas en sociedad, propiedad de hogares, su dirección está a cargo de hogares, en forma individual o en sociedad con otras personas. (Martinez, 2013)

Importancia de las pymes en el país.

- Aseguran el mercado de trabajo mediante la descentralización de la mano de obra, cumple un papel esencial en el correcto funcionamiento del mercado laboral.

- Tienen efectos socioeconómicos importantes ya que permiten la concentración de la renta y la capacidad productiva desde un número reducido de empresas hacia uno mayor.
- Reducen las relaciones sociales a términos personales más estrechos entre el empleador y el empleado favoreciendo las conexiones laborales ya que, en general, sus orígenes son unidades familiares.
- Presentan mayor adaptabilidad tecnológica y menor costo de infraestructura.
- Obtienen economía de escala a través de la cooperación entre empresas, sin tener que reunir la inversión en una sola firma.

Clasificación de la pequeña y mediana empresa según el Ministerio de Economía (MINEC):

a. Pequeña empresa

Según las consideraciones del banco multisectorial de inversiones, la pequeña empresa es aquella cuyas ventas anuales son $68,571.41 a $685,714.28, contando con un total de empleados entre 11 y 49. (MINEC, 2012)

b. Mediana empresa

Es la empresa cuyos ingresos oscilan entre los $685,714.28 hasta un total de $4,571,428.50, con un total de 50 a 199 empleados. (MINEC, 2012)

Las pymes son las empresas que mantienen un total de ventas anuales que oscilan entre los $68,571.28 hasta los $4,571,428.50 dólares y un total de entre 11 y 199 empleados. El Salvador tiene más de medio millón de micros, pequeñas y medianas empresas (pymes). Se calcula que emplean al 66 % de la población económicamente activa y aportan el 44 por ciento del Producto Interno Bruto (PIB). (MINEC, 2012)

En la Tabla 1, se presenta la clasificación en El Salvador de las pequeñas y medianas empresas.

Tabla 1. Clasificación de las pymes

Clasificación	Porcentaje de establecimiento (%)
Pequeña	7.54 %
Mediana	1.50 %
Total pymes	9.04 %

Fuente: sitio web de Cámara Salvadoreña de Comercio e Industria de El Salvador.

Distribución de las pymes en El Salvador

En El Salvador, actualmente existen un total de 13,208 empresas catalogadas como pequeña empresa y un total de 2,624 catalogadas como medianas empresas incluidas como pymes generan un total de 15,832 empresas cuya clasificación se encuentra dentro de pymes (Martínez, 2013). En cuanto a la distribución departamental de las pymes refleja la tendencia general de las firmas hacia la concentración geográfica, siendo los departamentos de San Salvador, San Miguel y Santa Ana los que concentran el 72.1 % del total de los establecimientos que emplean entre 5 y 99 empleados y, dentro de estos, la capital San Salvador reúne el 54% (Martínez, 2013)

En la gráfica mostrada en la Figura 22 se presenta la distribución de micro y mediana empresa.

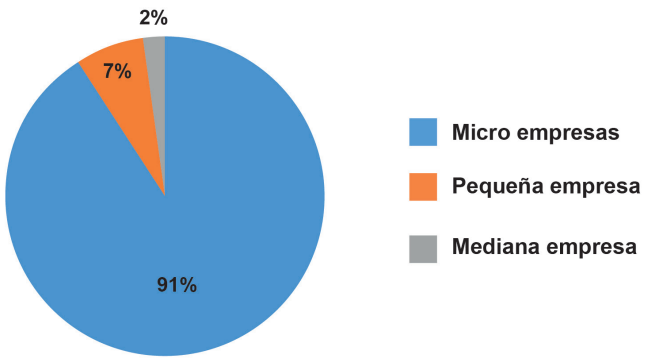


Figura 22. Distribución de las pymes

Fuente: sitio web de Cámara Salvadoreña de Comercio e Industria de El Salvador.

Los departamentos con mayor concentración de empresas pyme (San Salvador, San Miguel, Santa Ana y La Libertad) coincidirían, con excepción del departamento de Cuscatlán, con el grupo de departamentos que reportan un mayor nivel de ingreso y menor incidencia de pobreza. (Martínez, 2013)

Distribución económica de las pymes de acuerdo al sector en el cual se desempeñan

Existe un total de 17,5178 empresas en El Salvador clasificadas como pymes, de las cuales se hace una distribución generalizada por cada uno de los sectores, cuyo porcentaje está representado en la tabla 2.

Tabla 2. Cuadro de distribución por sector de las pymes

Sector	Porcentaje
Servicios	18.4 %
Transporte	2.3 %
Construcción	0.3 %
Electricidad	0.05 %
Agroindustria	0.04 %
Total	100 %

Fuente: sitio web de Cámara Salvadoreña de Comercio e Industria de El Salvador.

Los sectores que nos interesan estudiar son el de comercio y el de servicios.

Comercio electrónico

¿Qué es comercio electrónico?

El comercio electrónico consiste en realizar electrónicamente transacciones comerciales. Está basado en el tratamiento y transmisión electrónica de datos, incluidos texto, imágenes y vídeo. Por su parte, Guisado Moreno desde una perspectiva privatista, señala que se entiende por comercio electrónico aquel que abarca las transacciones comerciales electrónicas, compraventa de bienes y prestación de servicio realizados entre empresarios, o bien entre empresarios y consumidores, a través de los soportes electrónicos proporcionados por las nuevas tecnologías de la información y la comunicación− básicamente internet− así como las negociaciones previas y posteriores estrechas y directamente relacionadas con aquellos contratos (ofertas contractuales, contra ofertas, pago electrónico). (Gerencia, 2014)

Ventajas del comercio electrónico

La Web ofrece a los proveedores la oportunidad de relacionarse con un mercado totalmente interactivo, donde las transacciones, transferencias, inventarios y recolección de datos, entre otras actividades, pueden realizarse en línea. Esto permite que las empresas puedan incrementar su eficiencia, disminuyendo el tiempo de estas operaciones; automatizar los procesos de administración; acelerar la entrega de productos y mejorar la distribución. (Gerencia, 2014)

Entre otras ventajas, el comercio electrónico también permite que los consumidores cuenten con una plataforma de compra durante las 24 horas del día, y las empresas se introduzcan en un mercado focalizado a la medida de las necesidades de los clientes y al tiempo que disminuyen sus costos. Esta modalidad le brinda la oportunidad a las organizaciones de llegar a aquellos mercados geográficamente inalcanzables de manera rápida y eficaz, y entrar en un nuevo segmento de consumidores. (Gerencia, 2014)

Tipos de comercio electrónico

Algunas formas de hacer negocios electrónicamente se pueden listar:

a. Comercio electrónico B2B

El comercio electrónico B2B ***(Business to Business)*** es el negocio orientado entre las diversas empresas que operan a través de internet. (Gerencia, 2014)

Dentro del comercio electrónico B2B se pueden distinguir tres modalidades:

- El mercado controlado por el vendedor en busca de compradores.
- El mercado controlado por el comprador que busca proveedores.
- El mercado controlado por intermediarios que persiguen el acuerdo entre vendedores y compradores.

El comercio electrónico B2B ha supuesto un gran avance tecnológico, pero se requieren una serie de características para sacar el rendimiento óptimo. (Gerencia, 2014)

Experiencia en el mercado concreto.

- La oferta debe ser un valor añadido
- Evitar fallos de producción, logística y distribución

Las ventajas y características han convertido al comercio B2B en una opción que cada vez tiene más adeptos. (Gerencia, 2014)

- Reducción de costes
- Ampliación de mercado
- Aumento de la velocidad
- Centralización de oferta y demanda
- Información de compradores, vendedores, productos y precios en un lugar común
- Mayor control de las compras

b. Comercio electrónico B2C

En el comercio electrónico B2C (Business to Consumer) el negocio va dirigido de las empresas al consumidor. (Lozoya, 2013)

Las ventajas más destacables del comercio electrónico B2C son:

- Compras más cómodas y más rápida
- Ofertas y precios siempre actualizados
- Centro de atención al cliente integrado en la web

Los inconvenientes, como sucede en toda transacción, también existen. El consumidor debe prestar especial atención a la seguridad de compras. (Lozoya, 2013)

c. Comercio electrónico B2A

El comercio electrónico B2A (Business to Administration) es un servicio que ofrece la administración a las empresas –y también a los ciudadanos– para que se puedan realizar los trámites administrativos a través de Internet. (Lozoya, 2013)

Las ventajas para las empresas son evidentes:

- Ahorro considerable de tiempo y esfuerzo
- La posibilidad de descargarse formularios y modelos de los procedimientos administrativos
- Disponibilidad las 24 horas del día
- Información siempre actualizada

d. Comercio electrónico B2E

El comercio electrónico B2E (Business to Employee) es otra aplicación que, en este caso, relaciona a las empresas con sus empleados. A través de la intranet el empleado puede ejercer parte de sus funciones de los procesos de negocio de la empresa. (Lozoya, 2013)

El comercio electrónico B2E ofrece ventajas significativas:

- Menores costes y burocracia
- Formación en línea
- Mayor calidad en la información interna
- Equipos de colaboración en el entorno web
- Integración más ágil del profesional en la empresa
- Soporte para la gestión
- Comercio electrónico interno
- Fidelización del empleado

e. Comercio electrónico C2C

El comercio electrónico C2C (Consumer to Consumer) es el tipo de comercio que se lleva a cabo entre consumidores, bien sea mediante el intercambio de correos electrónicos o a través de tecnologías P2P (peer to peer). (Lozoya, 2013)

Una de las estrategias más comunes del comercio C2C para internet viene definida por aquel tipo de negocio cuyo objetivo es facilitar la comercialización de productos y/o servicios entre particulares. (Lozoya, 2013)

f. Comercio electrónico C2G

El comercio electrónico C2G (Citizen to Government) relaciona a los consumidores con el Gobierno, facilitando el intercambio telemático de transacciones entre los ciudadanos y las administraciones públicas. (Lozoya, 2013)

Algunos de los servicios más habituales son:

- Información
- Participación del ciudadano
- Suscripción para la notificación telemática
- Pago de tasas e impuestos
- Sugerencias y reclamaciones
- Entrada y/o salida a través de registro
- Diversos servicios, como empleo, sanidad o educación

g. Comercio electrónico B2G

El comercio electrónico B2G (Business to Government) busca una mejor optimización de los procesos de negociación entre empresas y el gobierno. Su aplicación se destina a los sitios o portales especializados en la administración pública. En ellos las instituciones oficiales tienen la posibilidad de contactar con sus proveedores, pudiendo estos agrupar ofertas o servicios. (Lozoya, 2013)

B. Marco teórico

Big Data

Big Data es un término aplicado a conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable. Se considera un conjunto de datos que crecen rápidamente y que no pueden ser manipulados por las herramientas de gestión de bases de datos tradicionales. (Aguilar, 2013)

El ser humano se ha visto en la necesidad de crear nuevas formas de comunicación y almacenar dicha información de manera constante siendo está de rápido crecimiento. Esta contribución a la acumulación masiva de datos se puede

encontrar en diversas industrias, las compañías mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores y operaciones, de la misma manera sucede con el sector público.

De acuerdo con un estudio realizado por ***Cisco Systems***, entre 2011 y 2016 la cantidad de tráfico de datos móviles crecerá a una tasa anual de 78 %, así como el número de dispositivos móviles conectados a Internet excederá el número de habitantes en el planeta. Las naciones unidas proyectan que la población mundial alcanzará los 7.5 billones para el 2016 de tal modo que habrá cerca de 18.9 billones de dispositivos conectados a la red a escala mundial, esto conllevaría a que el tráfico global de datos móviles alcance 10.8 Exabytes mensuales o 130 Exabytes anuales. (Rosa y Rivera Pleitez, 2016)

Este volumen de tráfico previsto para 2016 equivale a 33 billones de DVD anuales o 813 cuatrillones de mensajes de texto. Pero no solamente son los seres humanos quienes contribuyen a este crecimiento enorme de información, existe también la comunicación denominada máquina a máquina (M2M machine-to-machine) cuyo valor en la creación de grandes cantidades de datos también es muy importante. (TestingSoft, 2014)

En la siguiente infografía se observa una representación de Big Data:

En el cual describe la intervención del dato en todas las áreas de la Información y Tecnologías y cómo este se relaciona entre sí.

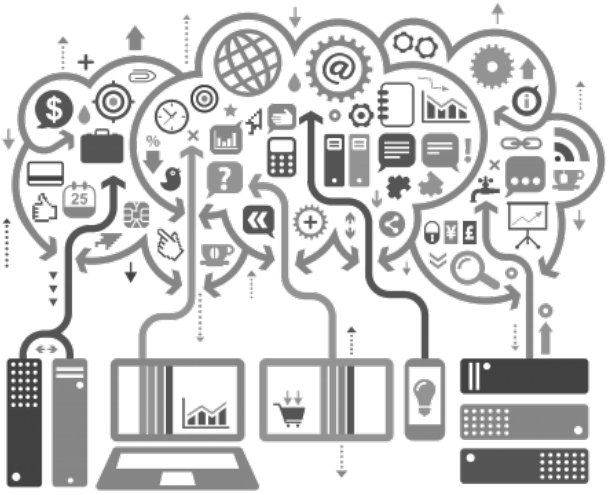


Figura 23. Infografía Big Data

Fuente: http://n-economia.com/notasalerta/trasformacion-digital-big-data-infografia/

Tipos de datos

- **Datos estructurados (Structured Data):** Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas.

Un ejemplo son las bases de datos relacionales y las hojas de cálculo.

- **Datos no estructurados (Unstructured Data):** Datos en el formato tal y como fueron recolectados, carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede desgranar su información a tipos básicos de datos. Algunos ejemplos son los PDF, documentos multimedia, e-mails o documentos de texto.

- **Datos semiestructurados (Semistructured Data):** Datos que no se limitan a campos determinados, pero que contiene marcadores para separar los diferentes elementos. Es una información poco regular como para ser gestionada de una forma estándar. Estos datos poseen sus propios metadatos semiestructurados que describen los objetos y las relaciones entre ellos, y pueden acabar siendo aceptados por convención. (Rosa y Rivera Pleitez, 2016)

Un ejemplo es el HTML, el XML, JSON, CSV.

Transformación

Una vez encontradas las fuentes de los datos necesarios, posiblemente dispongamos de un sinfín de tablas de origen sin estar relacionadas. El siguiente objetivo consta en hacer que los datos se recojan en un mismo lugar y darles un formato.

Aquí entran en juego las plataformas ETL (Extract, Transform and Load). Su propósito es extraer los datos de las diferentes fuentes y sistemas, para después hacer transformaciones (conversiones de datos, limpieza de datos sucios, cambios de formato…) y, finalmente, cargar los datos en la base de datos o Data Warehouse especificada. Un ejemplo de plataforma ETL es el Pentaho Data Integration, más concretamente su aplicación Spoon. (Rosa y Rivera Pleitez, 2016)

Almacenamiento NoSQL

El término NoSQL se refiere a Not Only SQL y son sistemas de almacenamiento que no cumplen con el esquema entidad-relación. Proveen un sistema de almacenamiento mucho más flexible y concurrente que permiten manipular grandes cantidades de información de manera mucho más rápida que las bases de datos relacionales. (Rosa y Rivera Pleitez, 2016)

Distinguimos cuatro grandes grupos de bases de datos NoSQL:

Almacenamiento clave-valor (Key-Value): Los datos se almacenan de forma similar a los maps o diccionarios de datos, donde se accede al dato a partir de una clave única. Los valores (datos) son aislados e independientes entre ellos, y no son interpretados por el sistema. Pueden ser variables simples como enteros o caracteres u objetos.

Por otro lado, este sistema de almacenamiento carece de una estructura de datos clara y establecida, por lo que no requiere un formateo de los datos muy estricto. Son útiles para operaciones simples basadas en las claves. Un ejemplo es el aumento de velocidad de carga de un sitio web que pueden utilizar diferentes perfiles de usuario, teniendo mapeados los archivos que hay que incluir según el id de usuario y que han sido calculados con anterioridad. (Rosa y Rivera Pleitez, 2016)

Almacenamiento documental: Las bases de datos documentales guardan un gran parecido con las bases de datos Clave-Valor, diferenciándose en el dato que guardan. Si en la anterior no requería una estructura de datos concreta, en este caso guardamos datos semiestructurados. Estos datos pasan a llamarse documentos, y pueden estar formateados en XML, JSON, Binary JSON o el que acepte la misma base de datos. Todos los documentos tienen una clave única con la que puede ser accedido e identificado explícitamente. (Rosa y Rivera Pleitez, 2016)

Almacenamiento en grafo: Las bases de datos en grafo rompen con la idea de tablas y se basan en la teoría de grafos, donde se establece que la información son los nodos y las relaciones entre la información son las aristas, algo similar en el modelo relacional. Su mayor uso se contempla en casos de relacionar grandes cantidades de datos que pueden ser muy variables. (Rosa y Rivera Pleitez, 2016)

Almacenamiento orientado a columnas: Por último, el almacenamiento Column-Oriented es parecido al documental. Su modelo de datos es definido como “un mapa de datos multidimensional poco denso, distribuido y persistente”. Se orienta a almacenar datos con tendencia a escalar horizontalmente por lo que permite guardar diferentes atributos y objetos bajo una misma clave. (Rosa y Rivera Pleitez, 2016)

¿Qué tipo de datos se deben explorar en Big Data?

Ya se ha mencionado que es tanta la información que se maneja hoy en día y existe mucha en la web; por lo tanto, dependerá de lo que se quiera analizar y el problema que se quiera resolver. Existe una gran variedad de datos y en distintos formatos, la clasificación de ellos se puede observar a continuación, aunque puede diversificarse de acuerdo a los avances tecnológicos:

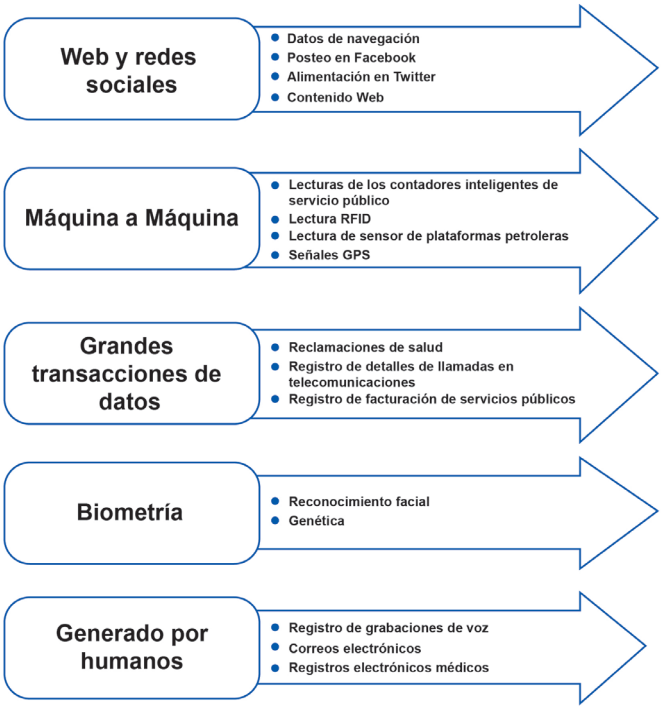


Figura 24. Clasificación de los datos Big Data

Fuente: Sitio web de IBM.

1. Web y redes sociales *(Web and Social Media)*: Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc., blogs.(IBM, developerworks, 2012)

2. Máquina a máquina (*Machine-to-Machine M2M)*: M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa. (IBM, developerworks, 2012)

3. Grandes transacciones de datos *(Big Transaction Data):* Incluye registros de facturación en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados. (IBM, developerworks, 2012)

4. Biometría (*Biometrics)*: Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación. (IBM, developerworks, 2012)

5. Generado por humanos *(Human Generated):* Las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc. (IBM, developerworks, 2012)

Componentes de una plataforma Big Data

Las empresas a nivel mundial han atacado esta problemática desde diferentes ángulos. Todas esas montañas de información generan un costo al no descubrir el valor asociado. Actualmente, quien tiene el liderazgo en términos de popularidad para analizar enormes cantidades de información es la plataforma de código abierto *Hadoop.*

Hadoop

Es utilizado en la actualidad por numerosas compañías para satisfacer sus necesidades de procesado de *big data*. Algunas de las grandes compañías que emplean Hadoop son Yahoo!, que lo emplea para realizar los cálculos requeridos por su motor de búsqueda o Facebook, que presume de tener el clúster más grande de Hadoop con más de 100 PB de datos.

Hadoop está inspirado en el proyecto de Google File System (GFS) y en el paradigma de programación *MapReduce*, el cual consiste en dividir en dos tareas (mapper – reducer) para manipular los datos distribuidos a nodos de un clúster logrando un alto paralelismo en el procesamiento. Hadoop está compuesto de tres piezas: *Hadoop Distributed File System* (HDFS), *Hadoop MapReduce* y *Hadoop Common.*

- **Hadoop Distributed File System(HDFS)**

Los datos en el clúster de Hadoop son divididos en pequeñas piezas llamadas *bloques* y distribuidas a través del clúster; de esta manera, las funciones `map` y `reduce` pueden ser ejecutadas en pequeños subconjuntos y esto provee de la escalabilidad necesaria para el procesamiento de grandes volúmenes.

HDFS es un sistema de ficheros que está especialmente diseñado para funcionar bien cuando se almacenan archivos grandes que, posteriormente, se leerán de forma secuencial. (Ghemawat S. G., 2003)

La siguiente figura ejemplifica como los bloques de datos son escritos hacia HDFS. Se observa que cada bloque es almacenado tres veces y al menos un bloque se almacena en un diferente rack para lograr redundancia.

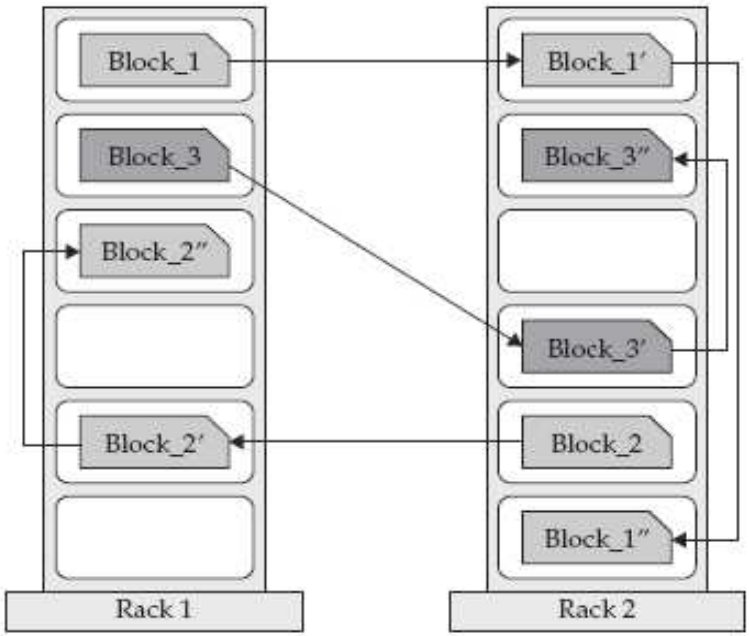


Figura 25. Ejemplo de DFS
Fuente: Sitio web IBM 2003.

- **Hadoop MapReduce**

El motor MapReduce es un sistema que gestiona los mecanismos para ejecutar tareas MapReduce de forma distribuida entre los diferentes nodos del clúster Hadoop. De nuevo, la forma en la que los datos se distribuyen en diferentes subtareas y cómo estas se asignan a cada máquina resulta transparente para el desarrollador. Además, el ecosistema de Hadoop se compone de otros proyectos que, sin ser vitales para su funcionamiento, permiten realizar determinadas tareas de un modo más sencillo o más eficiente. (Rosa y Rivera Pleitez, 2016)

La siguiente figura ejemplifica un proceso sencillo de MapReduce:

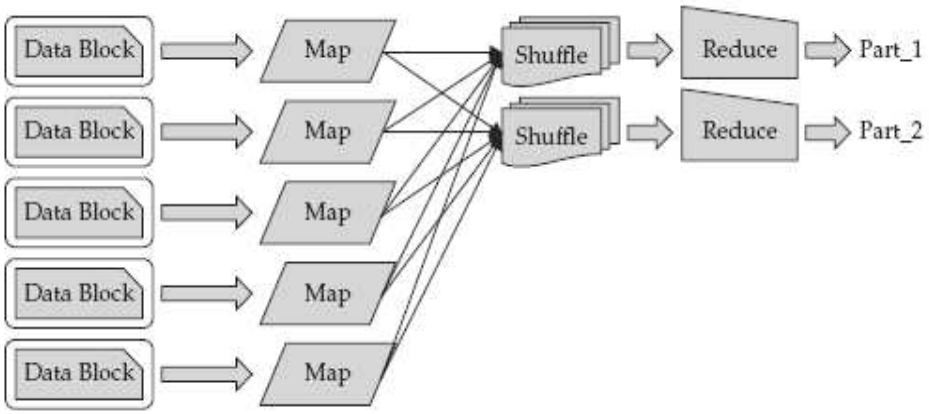


Figura 26. Ejemplo de MapReduce
Fuente: Sitio web IBM 2003.

- **Hadoop Common**

Hadoop Common Components son un conjunto de librerías que soportan varios subproyectos de Hadoop.

Principales distribuciones de Hadoop

En principio, todos los componentes del proyecto Hadoop; así como los demás proyectos relacionados, se pueden descargar del sitio web de Apache, donde se puede encontrar también documentación para llevar a cabo su instalación.

No obstante, muchas compañías han decidido lanzar sus propias distribuciones de Hadoop.

Las características de cada una de estas distribuciones dependen del fabricante, pero en general tienen las siguientes características:

- Ofrecen un ecosistema completo e interoperable. Como se comentó anteriormente, son muchos los pr oyectos que surgen en torno a Hadoop y, al evolucionar cada uno de ellos de forma independiente, en ocasiones hay que ser cuidadoso de escoger una combinación de versiones que funcione correctamente. Las distribuciones de Hadoop ofrecen un ecosistema

completo que ha sido testeado para garantizar que todos sus componentes funcionan correctamente.

- Ofrecen soporte (más allá del ofrecido por la comunidad de Hadoop), si bien este soporte no tiene por qué ser gratuito.

Además de las distribuciones (que pueden ser gratuitas o de pago), muchos fabricantes ponen a disposición de los usuarios lo que llaman un *sandbox*, en el que ofrecen una máquina virtual con Hadoop preinstalado (no apto para entornos en producción), que complementan con tutoriales u otros recursos útiles para los desarrolladores que no tengan experiencia previa.

a. Hortonworks

Una de las distribuciones más extendidas de Hadoop es Hortonworks Data Platform, que se presenta a sí misma como la distribución 100 % opensource de Hadoop y una de las que más ha contribuido al desarrollo de código del proyecto Hadoop.

Hortonworks incorpora numerosos proyectos que se integran con Hadoop para aumentar el abanico de posibilidades que ofrecer a los desarrolladores y a los usuarios. (Hortonworks Inc. 2016)

La siguiente figura muestra los diferentes componentes incluidos en Hortonworks, así como las versiones que se han escogido.

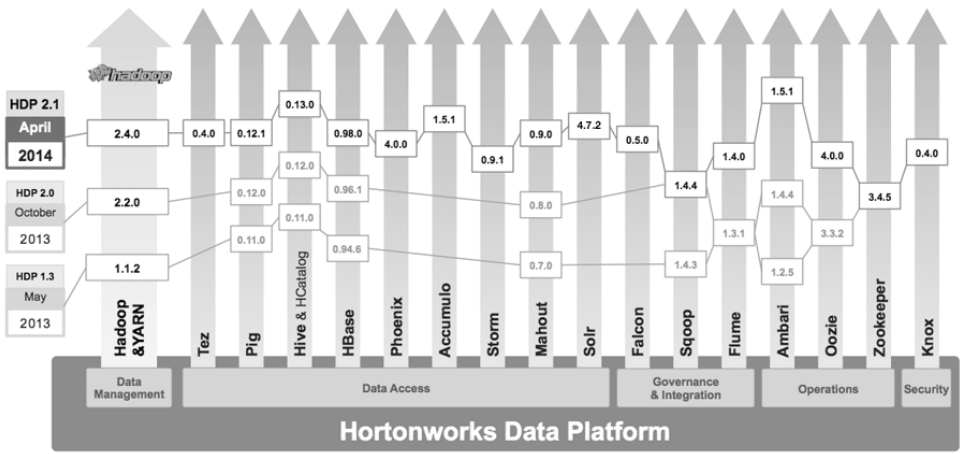


Figura 27. Componentes de Hortonworks Data Platform

Fuente: http://hortonworks.com/

Una de las particularidades de Hortonworks es que la distribución de Hadoop que ofrece está disponible tanto para sistemas GNU/Linux como para Microsoft Windows, siendo la única distribución a día de hoy que soporta este último sistema operativo. Además, Hortonworks ofrece una sandbox consistente en una máquina virtual con Hadoop preinstalado, complementado con el proyecto Hue (www.gethue.com), que ofrece una interfaz web sencilla y manejable para realizar algunas operaciones con Hadoop.

Además, Hortonworks ofrece la posibilidad de instalar y gestionar Hadoop a través de Ambari (ambari.apache.org), lo que simplifica sustancialmente la tarea de desplegar Hadoop en un clúster. (Rosa y Rivera Pleitez, 2016)

b. Cloudera

Cloudera ofrece CDH (Cloudera Data Hub) como distribución de Hadoop, que también se presenta como una distribución 100 % opensource al igual que Hortonworks. Según indica la propia empresa, esta distribución cuenta con un volumen de descargas que supera a la suma de todas las demás distribuciones juntas.

Basándose en esta distribución, Cloudera ofrece diferentes soluciones, tales por Cloudera Express (que es gratuita de forma ilimitada), o diferentes paquetes de Cloudera Enterprise, que ofrece funcionalidades complementarias y soporte adicional. Además de estas soluciones, también ponen a disposición de los usuarios Cloudera Live que permite probar Hadoop online utilizando la interfaz Hue, sin necesidad de instalarlo ni de descargar ninguna máquina virtual, por lo que resulta una alternativa interesante a Hortonworks Sandbox para usuarios que están empezando con Hadoop y no tienen recursos para desplegar una distribución. Finalmente, Cloudera es reconocido por ofrecer numerosos cursos de formación y también certificaciones en Hadoop, si bien estos cursos no son gratuitos. (Cloudera, Inc. 2016)

c. MapR

MapR ofrece M3 como distribución básica de Hadoop, disponible de forma gratuita para su descarga. Al igual que las otras distribuciones discutidas anteriormente, combina Hadoop con otros proyectos de Apache que le complementan y que ofrecen funcionalidad extra, además de una consola propia para la gestión del clúster.

Por encima de M3, MapR ofrece las distribuciones de M5 y M7, que añaden más funcionalidad para entornos en producción (tales como protección de datos, alta disponibilidad, etc.) y soporte extendido. Finalmente, MapR también ofrece la

posibilidad de descargar un sandbox para comenzar con Hadoop, funcionando sobre una máquina virtual que el usuario puede ejecutar en su ordenador. (MapR Technologies, Inc. 2016)

Investigación y ejemplos de aplicación de Big Data

En este apartado se hace referencia sobre investigaciones que se han hecho anteriormente, así como historia general sobre el auge de Big Data en la actualidad y como beneficiarse con las herramientas que existen en esta área.

Existen investigaciones sobre la aplicación de Big Data, entre ellas están las siguientes:

- Investigación de Big Data en los entornos de defensa y seguridad. (Carrillo Ruiz, y otros, 2013)
- El Big Data puede ayudar en el diagnóstico y tratamiento del cáncer. (Bernardo, 2013)
- Big Data, recurso para ciudades inteligentes (Souto, 2015) Big Data el nuevo recurso natural para las ciudades inteligentes el cual aprovecha múltiples fuentes de datos, este analiza los datos a través de la tecnología analítica y permite a los líderes servir mejor a los ciudadanos y negocios en un mundo cambiante. Este recurso aprovecha algoritmos predictivos para resolver los problemas proactivamente.
- Big Data ayuda al transporte inteligente en Dublín (Irlanda). Mediante el uso de herramientas de Big Data se puede ver el estado actual de toda la red de buses de un vistazo, y en forma detallada en áreas donde hay problemas para identificar la causa de la congestión nada más producirse y antes de que se extienda por otras rutas. La población de Dublín en el 2013 era de 1, 660,000. La información archivada ha servido para analizar a posteriori y entender lo que pasó y tomar medidas para optimizar el tráfico. (Zikopoulos, 2015)
- Big Data ayudó a Obama a ganar las elecciones el 2012 (Scherer, 2012). El equipo de dirección de campaña creó una mega base de datos con información de votantes y simpatizantes a partir de múltiples fuentes desde las elecciones de 2008. Analizaron la información a fin de identificar los gustos y preferencias de sus seguidores. Crearon diferentes aplicaciones para:

1. Mejores decisiones con mayor volumen de datos
2. Traducir datos en bruto para realizar análisis predictivo
3. Establecer las preferencias de los votantes
4. Es una solución a problemas de volumen

- Big Data ayuda a las tiendas Macy's de EEUU a incrementar sus ventas (De Juana, 2015). Hasta el 2010, Macy's seguía utilizando hojas de cálculo Excel para analizar grandes volúmenes de datos de clientes. Ahora, con Big Data, analiza decenas de millones de Terabytes (1012) de información cada día, y ha pasado de 22h a 19 min para rehacer el precio de sus artículos. Han conseguido:

  1. Mismas decisiones en menos tiempo
  2. Un incremento del 10% en sus ventas
  3. Tomar decisiones en menos tiempo

- Big Data ayuda a General Electric a mejorar sus productos (FondosFidelity, 2012). En el 2011, GE invirtió mil millones de dólares en un centro de investigación para mejorar sus diferentes productos. Allí analizan un gran volumen de datos procedentes de multitud de sensores y otros dispositivos digitales.

Las investigaciones citadas anteriormente solo son una pequeña muestra de las aplicaciones de Big Data en algunas áreas, pero en términos generales los principales sectores que usan esta tecnología son los siguientes:

- **Marketing** conocido también como Bussiness Intelligence o Inteligencia de Negocios, es muy utilizado en las empresas de muchos países del mundo, debido a que mediante estrategias bien definidas logran crear y administrar conocimiento sobre el medio, a través de análisis de los datos existentes dentro de la empresa.

- **Redes Sociales** son las que más beneficios han obtenido con el auge Big Data, pues mediante técnicas de recolección de datos, se puede llegar a saber las preferencias de las personas, lo cual es bien utilizado por las empresas para poder ser más competitivas.

- **Meteorología** maneja datos de gran tamaño y bases de datos de matemáticos desde hace mucho tiempo, por lo tanto, las herramientas Big Data han ayudado con ese problema.

- **Ingeniería** área muy amplia que se ve beneficiada con las tecnologías Big Data, entre ellas el transporte, sector energético y telecomunicaciones.
- **Medios de comunicación** debido a que están inmersos en todos los aspectos en el manejo de la información masiva. (Rosa y Rivera Pleitez, 2016)

Ventajas de utilizar herramientas de Big Data

El uso de las tecnologías Big Data en la actualidad ha beneficiado en gran manera el manejo de grandes volúmenes de datos, eso se ha podido comprobar en las investigaciones que se han mencionado anteriormente.

El uso de estas tecnologías permite el tratamiento y análisis de grandes repositorios de datos que, de otra manera, no sería posible si se quisiera lograr con las herramientas de bases de datos tradicionales, pues se vuelven insuficientes en todos los aspectos. Se ha mencionado que hoy en día una mayor cantidad de datos que se almacenan proceden de páginas web, aplicaciones de imágenes y videos, redes sociales, dispositivos móviles, apps o sensores; por lo tanto, es necesario contar con herramientas potentes que permitan almacenar, procesar y analizar esos datos para fines diversos: negocios, salud, comercialización de productos, etc. (Rosa y Rivera Pleitez, 2016)

Desventajas de utilizar herramientas de Big Data

Se mencionan muchos las ventajas que trae consigo el uso de las herramientas Big Data y se podría pensar que no hay desventajas que mencionar, aunque realmente si las hay, sobre todo, en países subdesarrollados en el que existe mucho desconocimiento en esta área, pero con enormes deseos por incursionar en las nuevas tecnologías.

Las desventajas que se pueden mencionar por la experiencia propia en El Salvador son las siguientes:

- Falta de profesionales expertos
- Resistencia al cambio por miedo al fracaso
- Falta de inversiones destinadas a implementar soluciones Big Dat
- Dificultad de integración en los procesos internos de las empresas
- Calidad de los datos

- Falta de interés por innovar y por capacitar al personal de las empresas

Por estas desventajas no podrían las empresas en El Salvador, y probablemente en otros países del mundo, decidirse a invertir para incursionar en la tecnología Big Data, pues la carencia de profesionales en esta área es bien limitada o nula. A parte que se debe invertir en equipo y capacitaciones y, por otro lado, hay que cambiar todos los procesos internos para migrar a las nuevas tecnologías. Muchas veces la resistencia al cambio y el miedo al fracaso truncan las posibilidades de mejorar y ser mucho más productivos profesionalmente. Sin embargo, para las empresas que han asumido ese reto los beneficios obtenidos de los procesos que se hacen en menor tiempo les permiten almacenar grandes volúmenes de datos. (Rosa y Rivera Pleitez, 2016)

En El Salvador, muchas empresas trabajan con las bases de datos relacionales tradicionales como SQL, Oracle, SyBase, MySQL que trabajan con datos que están estructurados. En el peor de los casos utilizan Excel o Access para el almacenamiento de los datos donde el tiempo de respuesta es mucho mayor y la cantidad de datos que se pueden almacenar no es en grandes volúmenes. Se sabe que los datos en la nube carecen de estructura y se almacena en formatos que no se pueden trabajar con las bases de datos relacionales, por lo cual se necesitan herramientas que permitan procesar y analizar grandes volúmenes de datos y en el menor tiempo posible. (Rosa y Rivera Pleitez, 2016)

A pesar de las desventajas que se mencionan, las empresas observan que cada día se encuentran con mayor cantidad de datos que provienen de fuentes diversas que muchas veces no proporcionan datos estructurados y es allí donde surge la necesidad de hacer uso de otras herramientas que no sean las convencionales.

Teorías

Big Data es un sistema genérico que debe tratar una gran cantidad de datos y que le hace falta integrar muchas herramientas para que sea lo que dice su nombre, dependiendo de la cantidad de datos, su tipo, relación entre los mismos, modelos y algoritmos a ejecutar.

En esencia, se trata de un conjunto de tecnologías y arquitecturas diseñadas para conseguir un mejor rendimiento de grandes volúmenes de información como ocurre con cualquier modelo de negocio, el factor clave para obtener beneficios de Big Data no depende de la capacidad tecnológica sino de la capacidad humana para realizar la correcta interpretación de la información que permita obtener valor de su análisis. (Rosa y Rivera Pleitez, 2016)

Potencial Big Data

Big Data no es solo una herramienta o una tecnología si no un conductor de una disciplina de toma de decisiones mejorada basada en análisis predictivos, que marca el comienzo de una era de cambio cultural y mejora del rendimiento. La experiencia del usuario será clave, no solo en la venta de servicios, sino también en los productos. Con Big Data la venta de productos o servicios podrá diferenciarse haciendo que el consumo de los mismos suponga una experiencia personalizada para los gustos y preferencias de cada cliente. Big Data permitirá llevar a cabo la gestión de emociones a la hora de enriquecer el consumo de los productos y servicios. (Rosa y Rivera Pleitez, 2016)

Big Data no es una actividad aislada. Para el éxito se necesita más que nunca el conocimiento del negocio que permita hacer las preguntas correctas y establecer las correlaciones oportunas. Negocio y TI deben de ir de la mano desde el primer momento y más que nunca. (Rosa y Rivera Pleitez, 2016)

Sin duda alguna, uno de los retos de Big Data es incorporar a su capacidad analítica, información de contexto que permita adaptar y comprender el resultado del análisis con base en las condiciones del entorno. Para ello, el verdadero conocimiento será aquel que incorpore los atributos de entorno que contextualicen el análisis. La contextualización del dato trata de responder e incorporar al análisis, información relativa a ¿cuándo se obtuvo la fuente origen?, ¿cómo se obtuvo?, ¿de dónde procede?, ¿Cuál es su naturaleza?

Existe una gran complejidad para realizar análisis cuando el número de variables es muy alto, mucha de la información puede no ser útil o considerarse falsa. Big Data puede derivar que se encuentren correlaciones falsas o falsos positivos. Para intentar solventar esta problemática en el rigor del análisis de los datos, existen ciertas premisas que ayudan a evitar errores, las más importantes son:

- Es primordial comenzar con pequeños pilotos para ganar experiencia y conocimiento de las nuevas tecnologías.
- Es recomendable trabajar con expertos para evitar cometer grandes errores.
- Construir un modelo que permita conocer a futuro y corrija los errores permitiendo la optimización de los procesos de negocio.

Un ejemplo lo podemos encontrar en Google, respecto a las predicciones sobre la epidemia de gripe en América del Norte, dos años más tarde el estudio ya no era válido debido a que los datos habían cambiado y el sistema no había sido realimentado de forma asistida. (Rosa y Rivera Pleitez, 2016)

Big Data es mucho más que volumen de información, muchos tipos de variables, muchos tipos de observaciones, muchos resultados, lo realmente importante es:

- Cómo están extraídos los datos
- De dónde provienen
- Su fiabilidad
- De todos los datos, cuáles son los que son relevantes a integrar en el sistema
- La relevancia forma parte de la respuesta
- Esto limita el alcance del sistema
- Influye sobre la definición misma del sistema

Por otro lado, para la meteorología, el tema ya está abordado completamente por la riqueza en cuanto a la gran cantidad de datos continuos con los que se alimenta el sistema. La evolución de orígenes de datos son constantemente reevaluadas con el fin de poder integrarlas de forma adecuada en el sistema. (Rosa y Rivera Pleitez, 2016)

A continuación una breve descripción de las herramientas que se van a utilizar en la metodología propuesta:

Descripción de las herramientas a utilizar en la investigación

¿Qué es Hadoop?

En la actualidad, Hadoop es un proyecto de *software* libre, con licencia Apache, cuya finalidad es prestar una plataforma para la gestión de grandes cantidades de datos. Los principales componentes que constituyen Hadoop son el sistema de archivos HDFS y el motor MapReduce.

HDFS (Hadoop Distributed File System) es un sistema de archivos distribuido inspirado en el GFS de Google que permite distribuir los datos entre distintos nodos de un clúster (llamados datanodes), gestionando la distribución y la redundancia de forma transparente para el desarrollador que vaya a hacer uso de esos datos. (Rosa y Rivera Pleitez, 2016)

El motor MapReduce es un sistema que gestiona los mecanismos para ejecutar tareas MapReduce de forma distribuida entre los diferentes nodos del clúster Hadoop. De nuevo, la forma en la que los datos se distribuyen en diferentes subtareas y cómo estas se asignan a cada máquina resulta transparente para el desarrollador. Además, el ecosistema de Hadoop se compone de otros proyectos

que, sin ser vitales para su funcionamiento, permiten realizar determinadas tareas de un modo más sencillo o más eficiente. Hadoop es utilizado en la actualidad por numerosas compañías para satisfacer sus necesidades de procesado de Big Data. Algunas de las grandes compañías que emplean Hadoop son Yahoo!, que lo emplea para realizar los cálculos requeridos por su motor de búsqueda, o Facebook, que presume de tener el clúster más grande de Hadoop con más de 100 PB de datos. (Rosa y Rivera Pleitez, 2016)

Existen varias distribuciones de Hadoop como Hortonworks, Cloudera, MapR para este trabajo de investigación se utilizará **Hortonworks.** Una de las distribuciones más extendidas de Hadoop es Hortonworks Data Platform que se presenta a sí misma como la distribución 100 % *opensource* de Hadoop y una de las que más ha contribuido al desarrollo de código del proyecto Hadoop.

Hortonworks incorpora numerosos proyectos que se integran con Hadoop para aumentar el abanico de posibilidades que ofrecer a los desarrolladores y a los usuarios. Una de las particularidades de Hortonworks es que la distribución de Hadoop que ofrece está disponible tanto para sistemas GNU/Linux como para Microsoft Windows, siendo la única distribución a día de hoy que soporta este último sistema operativo.

Además, Hortonworks ofrece una sandbox consistente en una máquina virtual con Hadoop preinstalado, complementado con el proyecto Hue (www.gethue.com), que ofrece una interfaz web sencilla y manejable para realizar algunas operaciones con Hadoop. Hortonworks ofrece la posibilidad de instalar y gestionar Hadoop a través de Ambari (ambari.apache.org), lo que simplifica sustancialmente la tarea de desplegar Hadoop en un clúster. (Rosa y Rivera Pleitez, 2016)

Para llevar a cabo el despliegue de Hadoop es necesario comenzar realizando los siguientes pasos:

1. Decidir la arquitectura física del sistema. Este paso es probablemente el más complicado en un principio, pues requiere tener una visión global del uso que se va a realizar del sistema. Esta decisión incluye conocer el número de nodos del clúster, los servicios que ejecutará cada uno de ellos, la distribución física de los equipos, etc. Una de las principales ventajas que ofrece Hadoop es que es relativamente sencillo adaptar la infraestructura física a las necesidades que surjan en un futuro (por ejemplo, añadiendo nuevos nodos)

2. Decidir la distribución que se va a desplegar. En caso de que se haga, como se ha comentado anteriormente, la distribución a utilizar será Hortonworks aunque todas las demás ofrecen combinaciones de diversos proyectos del ecosistema Hadoop cuya interoperabilidad ha sido testeada, además, de funcionalidades extra y soporte.

Para la elaboración de la propuesta metodológica se empleará una arquitectura basada en dos máquinas virtualizadas (de las cuales, una hará de máster y la otra de esclavo), esto por la carencia de recursos físicos de las cuales se dispone para llevar a cabo el despliegue.

En cuanto a la distribución se sugiere instalar Hortonworks, puesto que dispone de un instalador y configurador del clúster basado en Apache Ambari, lo que simplifica enormemente el proceso. (Rosa y Rivera Pleitez, 2016)

¿Por qué Hadoop?

En la actualidad, la cantidad de datos que se generan a cada segundo es inmensa. En el año 2012, IBM publicó una infografía en la que basándose en fuentes como estudios de IDC o EMC resumía el estado actual en lo referente a la cantidad de datos que inundaba la web. (International Business Machines Corp. 2012)

Algunas de las cifras más significativas son las siguientes:

- Se mandan 294.000 millones de emails diariamente
- Se suben 100 terabytes de datos a Facebook diariamente
- Se generan 5 exabytes (millones de terabytes) cada dos días
- Existen en el universo digital 2,7 zettabytes (miles de millones de terabytes) de datos

Esta inmensidad de datos se puede explicar fundamentalmente en base a tres orígenes distintos:

- La interacción entre humanos que emplean un sistema informático que registra información mientras se produce la interacción. Es el caso del correo electrónico, los foros de Internet o las redes sociales, en los que los datos los generamos los humanos y estos son almacenados o procesados por máquinas.
- La interacción entre un humano y una máquina. Este caso se da cuando navegamos por Internet y los servidores web generan logs con información sobre el proceso de navegación, o cuando compramos en una plataforma

de comercio electrónico o empleamos la banca online, y un sistema registra nuestras transacciones.

- La interacción entre máquinas (M2M), en las que son varias máquinas las que intercambian información entre ellas y registran esta información. Algunos ejemplos son sistemas de monitorización, en los que un sistema de sensores proporcionan la información que reciben a otras máquinas para que realicen algún procesado sobre ella. (Rosa y Rivera Pleitez, 2016)

- Probablemente, más importante que la cantidad de datos que se genera en la actualidad, es entender que este ritmo crecerá en el futuro, puesto que cada vez son más las personas que tienen acceso a Internet y la variedad de dispositivos que se conectan a la red. En esta misma infografía, IBM revela que en 2020 se generarán 35 zettabytes de datos anualmente.

Este ritmo de generación de datos introduce numerosos desafíos en lo que concierne al modo en que se almacenan estos datos y, especialmente, a la forma en la que deben ser procesados. Resulta evidente que los sistemas tradicionales son incapaces de manejar esta información de una forma eficiente, puesto que no están preparados para soportar la explosión de datos de los últimos años.

Las principales compañías que vinieron notando esta necesidad de sistemas para el almacenamiento y procesado más eficiente de grandes cantidades de datos fueron los buscadores de Internet.

La labor de un buscador de Internet es, en teoría, relativamente sencilla. Lo primero que debe hacer es rastrear la web, siguiendo los hipervínculos de cada página, para ir construyendo un grafo (una estructura de datos que relaciona diferentes nodos en este caso páginas web entre sí por medio de enlaces). A continuación, debe elaborar un índice invertido, donde se pueda localizar fácilmente una web dados unos términos de búsqueda. Además, los buscadores suelen emplear una función de ranking, por la que asignan a cada página web un peso en función de su relevancia, que se puede calcular en base a la cantidad de páginas que enlazan con esa página y, al mismo tiempo, a la relevancia de cada una de estas páginas. (Rosa y Rivera Pleitez, 2016)

La principal dificultad que encuentran los buscadores es que la cantidad de páginas web disponibles es inmensa, lo que dificulta llevar a cabo todo el proceso de indización en un tiempo razonable como para mantener los resultados de búsqueda actualizados. (Rosa y Rivera Pleitez, 2016)

En el año 2003 (Ghemawat, Gobioff, & Leung, 2003), Google publica su famoso artículo en el que describe Google File System, un sistema de ficheros escalable y distribuido que pretende subsanar la dificultad de tener que almacenar grandes

cantidades de datos de forma confiable y proporcionando un alto rendimiento para aplicaciones que realizan un uso intensivo de estos datos.

Al año siguiente, en 2004, Google publica otro artículo (Dean, MapReduce: Simplified Data Processing on Large, 2004) en el que describe el paradigma de programación MapReduce, cuya finalidad es llevar a cabo un procesado distribuido de grandes cantidades de datos de forma eficiente.

MapReduce

MapReduce es un desarrollo que responde a la necesidad de Google de procesar grandes cantidades de datos de manera eficiente, de forma paralela. Además, es un paso intuitivo tras el desarrollo de Google File System (GFS): puesto que ahora hay grandes cantidades de datos almacenadas de forma distribuida entre varios equipos, resulta oportuno realizar un procesamiento también distribuido de estos datos. (Rosa y Rivera Pleitez, 2016)

La idea detrás de MapReduce es sencilla: una aplicación MapReduce cuenta con una rutina ***map()*** y otra rutina ***reduce()***, que son las que dan nombre a este modelo de programación. La rutina ***map()*** recibe una tupla clave-valor ***(<k, v>)*** y devuelve un conjunto de tuplas clave-valor ***(<km, vm>).*** (Rosa y Rivera Pleitez, 2016)

Posteriormente, todas las claves devueltas por las rutinas ***map()*** ejecutadas se ordenan y se agrupan por clave, resultando un conjunto de tuplas que contienen una clave y una lista de valores. Por ejemplo, si las rutinas ***map()*** habían devuelto las tuplas ***<kmi, vmi,1>, <kmi, vmi,2>, …, <kmi, vmi,n>,*** tras esta fase todas estas tuplas se agruparán en una tupla ***<kmi, [vmi,1, vmi,2, …, vmi,n]>.***

Por último, la rutina reduce() recibe como entrada estas tuplas agrupadas y devuelven, para cada una de ellas, un conjunto de valores ***(<vr>).***

Probablemente trabajar bajo este paradigma de programación, para muchos no sea fácil de comprender y aplicarlo, pero por esa razón se trabajará con una herramienta que viene incluida dentro de Hadoop, la cual es mucho más fácil, sobre todo para aquellos que en alguna ocasión han trabajado con base de datos relacionales como SQL. Esta herramienta es conocida como Hive, la cual incorpora el procesado MapReduce.

Hive

Apache Hive es un proyecto que forma parte del ecosistema Hadoop y, por ello, viene incluido en muchas distribuciones de Hadoop, incluyendo la distribución Hortonworks.

El propósito de Hive es, en cierto modo, emular un sistema de bases de datos relacional encima de Hadoop.

Así, el usuario podrá crear tablas e insertar datos (o crearlas a partir de ficheros existentes en HDFS), para posteriormente consultarlas empleando un lenguaje de modelado y de consulta muy similar a SQL.

Es importante entender que esta lógica funciona bien cuando trabajamos con datos que son estructurados, puesto que el concepto de tablas en el modelo relacional estructura los datos en columnas (campos) y en filas (registros).

Hive, es una herramienta adecuada para usuarios que estén familiarizados las bases de datos relacionales. Permite crear tablas y hacer consultas sobre ellas empleando un lenguaje similar a SQL, si bien estas consultas se traducirán automáticamente a rutinas MapReduce. (The Apache Software Foundation, 2014)

¿Qué es R?

Se puede definir R desde dos perspectivas distintas:

- R es un entorno de software
- R es un lenguaje de programación

Fundamentalmente R puede ser definido como un entorno software para el análisis matemático y estadístico de datos, en cierto sentido similar a herramientas tales como Microsoft Excel. A través del entorno de R vamos a ser capaces de manipular datos (por ejemplo, cargarlos desde ficheros, editarlos, volverlos a almacenar...), realizar análisis sobre esos datos y presentar los resultados gráficamente para facilitar su interpretación.

El entorno software viene acompañado de un lenguaje de programación que pone a nuestra disposición las funcionalidades típicas de un lenguaje de propósito general (manejo de variables, tipos y estructuras de datos, operadores, mecanismos de control del flujo de ejecución, funciones, etc.) combinadas con librerías y herramientas específicas para facilitar el análisis de datos. Utilizando este lenguaje es relativamente sencillo implementar nuestras propias funciones y scripts para automatizar el procesamiento de ciertos datos.

En la práctica, estas dos perspectivas están muy relacionadas, así por ejemplo para interactuar con el entorno de R se utilizarán expresiones escritas en el lenguaje R.

¿Por qué utilizar R?

Actualmente existe una amplia gama de herramientas que pudiéramos pensar en utilizar a la hora de llevar a cabo análisis de datos (como por ejemplo Microsoft Excel, S-PLUS una versión comercial del lenguaje S, SAS, SPSS de IBM, etc.). Así pues, una de las preguntas que las empresas pudieran plantear en esta metodología es por qué elegir R como herramienta de análisis de datos.

Algunas de las razones que podría emplear a la hora de justificar la decisión incluyen:

- R es software de código libre con licencia GNU GPL (General Public License).
- Mientras que las principales herramienta de análisis son de pago (algunas con precios bastante elevados), R es completamente gratuito.
- Existen versiones para los sistemas operativos más comunes: Windows, Mac OS X y Linux.
- Posee una comunidad de usuarios amplia y muy activa, con lo que va a resultar relativamente sencillo encontrar documentación o ayuda en foros si resulta necesario.
- El entorno es fácilmente extensible, mediante el desarrollo de paquetes. Debido a esto, evoluciona rápidamente: nuevos algoritmos y técnicas de análisis se incorporan con regularidad.
- R y sus extensiones nos ofrece una gran variedad de herramientas de análisis y visualización de datos. Actualmente existen más de 5000 paquetes disponibles para ser instalados en el entorno. (The R Foundation, 2016)

Herramientas de visualización

Google Chart: es una aplicación de Google para realizar estadísticas web, de fácil uso para desarrolladores de software web, usado en muchos campos como Google Analytics, se puede usar con diferentes formatos, Json, Javascript y plugins que se pueden integrar con varios lenguajes de programación.

Esta herramienta, permite realizar gráficos atractivos y existe una gran variedad de gallerías disponibles en el sitio de Google para poder utilizarlos y adaptarlos a las necesidades de análisis de cada persona. (Google Chart. Inc, 2016)

Jqplot: es un framework para el trazado de gráficos y plugin jQuery Javascript, jqPlot. Produce hermosas líneas, barras y gráficos circulares con muchas características:

- Numerosas opciones de estilo gráfico
- Fecha ejes con formato personalizable
- Hasta 9 ejes Y
- Texto eje girado
- Cálculo automático de la línea de tendencia
- La información sobre herramientas y punto de datos resaltado
- Valores predeterminados razonables para facilitar su uso. (Google Chart .Inc, 2016)

D3.js: o simplemente D3 de documentos basados en datos, es una biblioteca JavaScript para producir visualizaciones de datos dinámicos e interactivos en los navegadores web. Hace uso ampliamente de SVG, HTML5 y estándares CSS. En contraste con muchas otras bibliotecas, D3.js permite un gran control sobre el resultado visual final.

Para poder hacer uso de esta herramienta, es necesario conocer de JavaScript, por consiguiente es necesario aprender ese lenguaje de programación. (Google Chart .Inc, 2016)

C. Marco conceptual

En definición, una plataforma es un sistema que sirve como base para hacer funcionar determinados módulos de hardware o de software con los que sea compatibles.

La Big Data genera enormes cantidades de información lo cual a su vez generan un costo potencial que no pueden solventar plataformas comunes. Por lo que es indispensable, obtener una que se adecue a analizar enormes cantidades de información, en ese sentido, una de las más conocidas es Hadoop. (IBM, 2012)

Algunas de las tecnologías asociadas a Hadoop

Hadoop

Hadoop es una plataforma de código abierto, está inspirado en el proyecto de Google File System (GFS) y en el paradigma de programación MapReduce, el cual consiste en dividir en dos tareas (Mapper–Reducer) para manipular los datos distribuidos a nodos de un clúster logrando un alto para el mismo en el procesamiento. Hadoop se compone de tres piezas. (International Business Machines Corp, 2012)

Hadoop Distributed File System (HDFS)

Los datos en el clúster de Hadoop son divididos en pequeñas piezas llamadas bloques y distribuidas a través del clúster; de esta manera, las funciones Map y Reduce pueden ser ejecutadas en pequeños subconjuntos y esto provee de la escalabilidad necesaria para el procesamiento de grandes volúmenes. (International Business Machines Corp, 2012)

A continuación se ejemplifica el flujo de HDFS como se observa en la figura 28.

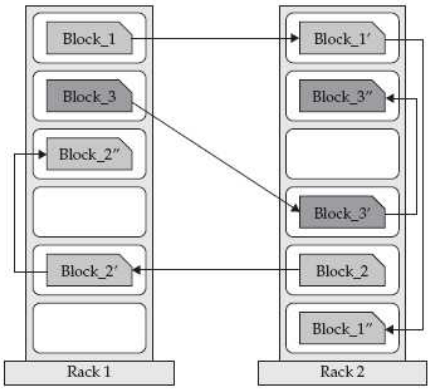


Figura 28. Ejemplo de HDFS

Fuente: Sitio Web de IBM.

Hadoop MapReduce

MapReduce es el núcleo de Hadoop. El término MapReduce en realidad se refiere a dos procesos separados que Hadoop ejecuta, el primer proceso Map toma un conjunto de datos y lo convierte en otro conjunto, donde los elementos individuales son separados en tuplas donde una tupla es una secuencia ordenada de objetos (pares de llave/valor). El proceso Reduce obtiene la salida de Map como datos de entrada y combina las tuplas en un conjunto más pequeño de las mismas. Una fase intermedia es la denominada Shuffle la cual obtiene las tuplas del proceso Map y determina que nodo procesará estos datos dirigiendo la salida a una tarea, reduce en específico tal como se muestra en la figura 29. (Zoho, Corp.2016)

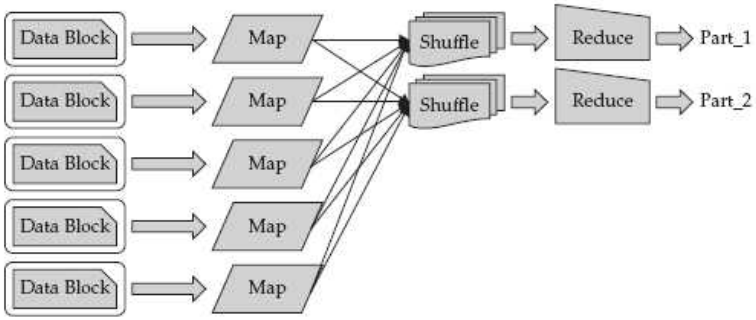


Figura 29. Ejemplo de ManReduce
Fuente: Sitio Web de IBM.

a. Hadoop Common

Hadoop Common Components son un conjunto de librerías que soportan varios sub proyectos de Hadoop. (International Business Machines Corp, 2012)

Además de estos tres componentes principales de Hadoop existen otros proyectos relacionados los cuales son definidos a continuación:

1. Avro

Es un proyecto de Apache que provee servicios que se serializarían, cuando se guardan datos en un archivo, el esquema que define ese archivo es guardado dentro del mismo; de este modo, es más sencillo para cualquier aplicación leerlo posteriormente puesto que el esquema está definido dentro del archivo. (International Business Machines Corp, 2012)

2. Cassandra

Es una base de datos no relacional distribuida y basada en un modelo de almacenamiento de <clave-valor>, desarrollada en Java. Permite grandes volúmenes de datos en forma distribuida. Twitter es una de las empresas que utiliza Cassandra dentro de su plataforma. (Aguilar, 2013)

3. Chukwa

Diseñado para la colección y análisis a gran escala de “logs”. Un log es un registro oficial de eventos durante un rango de tiempo en particular. Incluye un toolkit (conjunto de herramientas) para desplegar los resultados del análisis y monitoreo. (Aguilar, 2013)

4. Flume

Tal como su nombre lo indica, su tarea principal es dirigir los datos de una fuente hacia alguna otra localidad, en este caso hacia el ambiente de Hadoop. Las entidades principales en Flume. (Aguilar, 2013)

i Sources: Es cualquier fuente de datos.

ii Decorators: Es una operación dentro del flujo de datos que transforma esa información de alguna manera, como por ejemplo comprimir o descomprimir los datos o alguna otra operación en particular sobre los mismos. (Aguilar, 2013) iv. Sinks: Es el destino de una operación en específico.

5. HBase

Es una base de datos en columnas (column-oriented data base) que se ejecuta en HDFS. HBase no soporta SQL, de hecho, HBase no es una base de datos relacional. Cada tabla contiene filas y columnas como una base de datos relacional. HBase permite que muchos atributos sean agrupados llamándolos familias de columnas, de tal manera que los elementos de una familia de columnas son almacenados en un solo conjunto. (International Business Machines Corp, 2012)

6. Hive

Es una infraestructura de data warehouse que facilita administrar grandes conjuntos de datos que se encuentran almacenados en un ambiente distribuido. Hive tiene definido un lenguaje similar a SQL llamado Hive Query Language (HQL), estas sentencias HQL son separadas por un servicio de Hive y son enviadas a procesos MapReduce ejecutados en el cluster de Hadoop. (Aguilar, 2013)

7. Jaql

Fue donado por IBM a la comunidad de software libre. Query Language for Javascript Object Notation (JSON) es un lenguaje funcional y declarativo que permite la explotación de datos en formato JSON diseñado para procesar grandes volúmenes de información. Para explotar el paralelismo, Jaql reescribe los queries de alto nivel (Cuando es necesario) en queries de “bajo nivel” para distribuirlos como procesos. (Aguilar, 2013)

MapReduce. Internamente el motor de Jaql transforma el query en procesos Map y reduce para reducir el tiempo de desarrollo asociado en analizar los datos en Hadoop.

Jaql posee de una infraestructura flexible para administrar y analizar datos semiestructurados como XML, archivos CSV, archivos planos, datos relacionales. (Rivas, 2012)

8. Lucene

Es un proyecto de Apache bastante popular para realizar búsquedas sobre textos. Lucene provee de librerías para indexación y búsqueda de texto que ha sido, principalmente, utilizado en la implementación de motores de búsqueda (aunque hay que considerar que no tiene funciones de "crawling" ni análisis de documentos HTML ya incorporadas). (Rivas, 2012)

A nivel de arquitectura, Lucene es simple, básicamente los documentos (document) son divididos en campos de texto (fields) y se genera un índice sobre estos campos de texto. La indexación es el componente clave de Lucene, gracias a esto le permite realizar búsquedas rápidamente independientemente del formato del archivo sean PDF, documentos HTML. (Pontigo, 2013)

9. Oozie

Este fue un proyecto de código abierto que simplifica los flujos de trabajo y la coordinación entre cada uno de los procesos. Permite que el usuario pueda definir acciones y las dependencias entre dichas acciones. (Pontigo, 2013)

Un flujo de trabajo en Oozie es definido mediante un grafo a cíclico llamado Directed Acyclical Graph (DAG), y es a cíclico puesto que no permite ciclos en el grafo; es decir, solo hay un punto de entrada y de salida y todas las tareas y dependencias parten del punto inicial al punto final sin puntos de retorno. (Pontigo, 2013)

10. Pig

Inicialmente desarrollado por Yahoo para permitir a los usuarios de Hadoop enfocarse más en analizar todos los conjuntos de datos y dedicar menos tiempo en construir los programas MapReduce. (Pontigo, 2013)

Tal como su nombre lo indica al igual que cualquier cerdo que come cualquier cosa, el lenguaje PigLatin fue diseñado para manejar cualquier tipo de dato y Pig

es el ambiente de ejecución donde estos programas son ejecutados, de manera muy similar a la relación entre la máquina virtual de Java (JVM) y una aplicación Java. (Pontigo, 2013)

11. ZooKeeper

Es otro proyecto de código abierto de Apache que provee de una infraestructura centralizada y de servicios que pueden ser utilizados por aplicaciones para asegurarse de que los procesos a través de un clúster sean serializados o sincronizados. (Olguín, 2013)

El mantenimiento de la información de tipo de estado se realiza a través de la memoria en los servidores ZooKeeper, este es una máquina que mantiene una copia del estado de todo el sistema y persiste esta información en los archivos de registro locales. (Olguín, 2013)

Un gran grupo de Hadoop puede ser apoyado por varios servidores Zookeeper (en este caso un servidor maestro sincroniza los servidores de nivel superior) cada máquina cliente se comunica con uno de los servidores de ZooKeeper para recuperar y actualizar su información de sincronización. Dentro de un servidor, una aplicación puede crear lo que se llama un znode (un archivo que persiste en la memoria en los servidores de ZooKeeper). (Olguín, 2013)

El znode se puede actualizar todos los nodos del clúster y cualquier nodo del clúster puede registrarse para ser informado de los cambios a la znode (en ZooKeeper jerga, un servidor puede ser configurado para «vigilar» a znode específico). Con esta infraestructura znode (y hay mucho más a esto de tal manera que no podemos ni siquiera empezar a hacer justicia en esta sección) las aplicaciones pueden sincronizar sus tareas en el clúster distribuido mediante la actualización de su estado en un ZooKeeper znode, lo que haría a continuación informar al resto del grupo de cambio de estado de un nodo específico. (Olguín, 2013)

Herramientas de Código Abierto para el tratamiento de Big Data

La Big Data no es solo uno de los nuevos términos de moda, además se ha vuelto una de las grandes tendencias de los últimos años, al hablar de la Big Data hablamos de “oro digital” dentro de la era de la información, pero el termino Big Data no se debe a los datos en sí, sino al crecimiento exponencial que estos están sufriendo gracias al uso de las nuevas tecnologías, es por esto que se han desarrollado sistema o herramientas que permitan obtener el mejor conocimiento de estas grandes cantidades de información. (Aguilar, 2013)

A continuación se muestra la figura 30 una rápida evolución en los medios de almacenamientos:

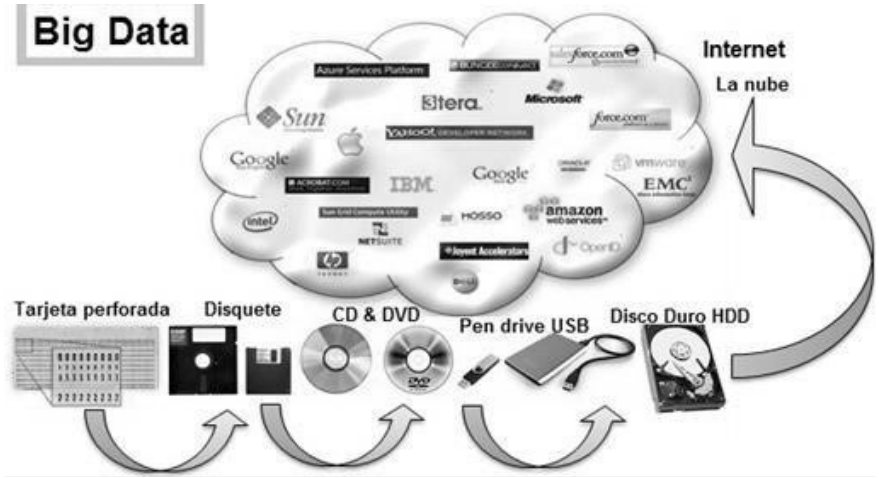


Figura 30. Evolución de los medios de almacenamiento.

Fuente: revista Cloud de almacenamiento.

Algunas de las herramientas de análisis para la Big Data mayormente utilizadas hoy en día en el mundo código abierto (Open Source) las mencionamos a continuación:

a. Herramientas framework para Big Data

i. MapReduce

Es un framework (modelo de programación) que hace honor a la frase célebre de Julio César "divide y vencerás", y que permite utilizar el procesamiento paralelo de datos en computadoras distribuidas. (Int ernational Business Machines Corp, 2013)

ii. Hadoop

Simplemente no se habla de Big Data sin mencionar a Hadoop ya que permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. Storm conocido como el Hadoop en tiempo real. (International Business Machines Corp, 2013)

iii. Pig/PigLatin

Un proyecto del ecosistema de Big Data de la fundación apache que utiliza un lenguaje textual el PigLatin con la plataforma de análisis Pig. (International Business Machines Corp, 2013)

b. Herramientas para Bases de Datos y DataWarehouse

i **HBASE:** Es el sistema de almacenamiento no relacional para Hadoop.

ii **CASSANDRA:** Anteriormente menciona es otro sistema de almacenamiento NoSQL desarrollado originalmente por Facebook.

iii **MongoDB:** Su nombre de la palabra en inglés "humongous" que significa enorme) y es un sistema de base de datos que es par te de la familia NoSQL, orientado a documentos, desarrollado bajo el concepto de código abierto. (Oracle, 2013)

iv **Neo4j:** Software libre de bases de datos orientado a grafos, los desarrolladores describen a Neo4j como un motor de persistencia embebido, basado en disco, completamente transaccional Java que almacena datos estructurados en grafos más que en tablas. (Oracle, 2013)

v **Ryak:** Declara ser el mejor sistema en producción de bases de datos distribuida Open Source. Sus características principales son: escalabilidad, tolerancia a fallos, alta disponibilidad y replicación. (International Business Machines Corp, 2013)

vi **HyPertable:** Este gesto de bases de datos es un sistema de almacenamiento de datos distribuidos y de alto desempeño NoSQL, ideal para aplicaciones que necesitan manejar datos que evolución rápidamente y soporta una gran demanda de datos en tiempo real, basándose en el diseño de BigTable de Google Inc. (IBM, 2013)

vii **Hive:** es el Datawarehouse de Hadoop, ya que es un sistema de almacenamiento que facilita el resumen de datos fácil, consultas ad-hoc y el análisis de grandes conjuntos de datos almacenados en Hadoop. (Lozano, 2014)

viii **Redis:** es un motor de base de datos en memoria, basado en el almacenamiento entablas de hashes (clave/valor) pero que opcionalmente puede ser usada como una base de datos durable o persistente patrocinado por VMware V. (International Business Machines Corp, 2013)

c. Herramientas para inteligencia de negocios

i Pentaho

Una de las herramientas más utilizadas para el análisis de la información orientada a la toma de decisiones empresariales, incluye algunas herramientas integradas para generar informes, perteneciendo a la comunidad Open Source. (Bustillo, 2013)

ii Palo BI Suite/Jedox

Una suite completa para la administración de un DataWarehouse cuenta con todas las aplicaciones básicas para crear solución para toma de decisiones BI de alta gama que está disponible libre de cuotas por licencias. (Jedox, 2013)

iii Jaspersoft

Es una de las herramientas más completas que tiene el Open Source que aloja sus proyectos de la comunidad en jasperforge.org, cuenta con funciones integradas de informes, dashboards, análisis e integración de datos diseñado para ayudar a las empresas a la toma de decisiones más rápidas y acertadas. (International Business Machines Corp, 2013)

iv Talend

Tiene una suite de herramientas Talend Open Studio for Big Data la cual se integra con el ecosistema Hadoop. (Stratebi, 2014)

d. Herramientas para la minería de datos

i Mahout

Un proyecto de la fundación apache para producir implementaciones libres de algoritmo de aprendizaje automático distribuido o escalables

enfocados en las áreas de filtrado colaborativo, agrupación y clasificación que hace parte del ecosistema hadoop y que pretende ser un sistema escalable de máquinas de aprendizaje. (International Business Machines Corp, 2013)

ii SAS Enterprise Miner / SAS

Solución de minería de datos que proporciona gran cantidad de modelos y de alternativas. Permite determinar pautas y tendencias, explica resultados conocidos e identifica factores que permiten asegurar efectos deseados. Además, compara los resultados de las distintas técnicas de modelado, tanto en términos estadísticos como de negocio, dentro de un marco sencillo y fácil de interpretar. (Analytics Software & Solutions, 2014)

iii Clementine / SPSS

Herramienta de data mining que permite desarrollar modelos predictivos y desplegarlos para mejorar la toma de decisiones. Está diseñada teniendo en cuenta a los usuarios empresariales de manera que no es preciso ser un experto en data mining. (International Business Machines Corp, 2013)

iv RapidMiner/RapidAnalytics

Herramientas con completo análisis de banco de trabajo y con fuerte para centrarse en la minería de datos y análisis predictivo. RapidMiner cubre un amplio rango de minería de datos. (International Business Machines Corp, 2013)

Además de ser una herramienta flexible para aprender y explorar la minería de datos, la interfaz gráfica de usuario tiene como objetivo simplificar el uso para las tareas complejas de esta área. (International Business Machines Corp, 2013)

v SAS Analytics / SAS

Suite de soluciones analíticas que permiten transformar todos los datos de la organización en conocimiento, reduciendo la incertidumbre, realizando predicciones fiables y optimizando el desempeño. (Analytics Software & Solutions, 2014)

El internet de las cosas

a. ¿Qué es el internet de las cosas?

Sus siglas en inglés (Internet of things) se refiere al internet de las cosas a una red de objetos cotidianos interconectado, surgió como una idea que se basa en que exista una capa de conectividad digital para cosas existentes, donde «cosas» se refiere a todo tipo de objetos de uso diario e incluso a sus componentes. Se espera que esta idea traiga consigo beneficios a corto plazo en aspectos como: optimización de la cadena de abastecimiento, efectividad de costos, mejoras en las experiencias de los consumidores, y beneficios en aspectos de seguridad y servicios de emergencia. (Castro, 2014)

No existe una definición unificada sobre qué es el internet de las cosas. La mejor forma de explicar el concepto es si te imaginas un mundo en el que todos los objetos cotidianos tienen una relación o una referencia digital, mediante una tecnología similar a RFID. En este mundo todos los objetos y sus partes estarían contabilizados, haciendo prácticamente imposible que un objeto se pierda o carezca de componentes; además, de que los objetos se pueden comunicar entre sí, almacenando e intercambiando información. (Castro, 2014)

Por supuesto, para hacer realidad una idea como el internet de las cosas, se requiere una evolución en la tecnología que soporta a internet (por ejemplo, el volumen de información siendo intercambiado simultáneamente), además de cambios en el paradigma de lo cotidiano. (Castro, 2014)

Imagina un refrigerador que hace un inventario de su contenido con la capacidad de hacer una lista para ir al supermercado, y de enviar esta lista a tu dispositivo móvil. Esto no está lejano de la realidad; ya se está hablando de los smartrefrigerators, e incluso hay algunos que ya están en el mercado, con aplicaciones disponibles y conexión a Wi-Fi. En general, ya se habla del término smartappliances para referirse a este tipo de electrodomésticos que también incluyen smart oven, smartlaundry y smartvaccum. (Castro, 2014)

A continuación la figura 31 muestra una pequeña representación del internet de las cosas.

Figura 31. Representación del internet de las cosas.

Fuente: http://www.cromo.com.uy

b. Seguridad de la información en la web

i. De qué trata la seguridad de la información

La seguridad de la información es el conjunto de medidas preventivas y reactivas que utilizan las organizaciones y de los sistemas tecnológicos que les permitan resguardar y proteger la información buscando mantener siempre el más alto grado de confidencialidad, la disponibilidad e integridad de la misma. (Castro, 2014)

El concepto de seguridad de la información no debe ser confundido con el de seguridad informática, ya que este último solo se encarga de la seguridad en el medio informático, pero la información puede encontrarse en diferentes medios o formas, y no solo en medios informáticos. (Castro, 2014)

Para el hombre como individuo, la seguridad de la información tiene un efecto significativo respecto a su privacidad, la que puede cobrar distintas dimensiones dependiendo de la cultura del mismo.

El campo de la seguridad de la información ha crecido y evolucionado considerablemente a partir de la Segunda Guerra Mundial, convirtiéndose en una carrera acreditada a nivel mundial. (Castro, 2014)

Este campo ofrece muchas áreas de especialización, incluidos la auditoría de sistemas de información, planificación de la continuidad del negocio, ciencia forense digital y administración de sistemas de gestión de seguridad, entre otros. (Castro, 2014)

La información es poder, y según las posibilidades estratégicas que ofrece tener acceso a cierta información, esta se clasifica como:

- Crítica: Es indispensable para la operación de la empresa
- Valiosa: Es un activo de la empresa y muy valioso
- Sensible: Debe de ser conocida por las personas autorizadas

Existen dos palabras muy importantes que son riesgo y seguridad:

- Riesgo: Es la materialización de vulnerabilidades identificadas, asociadas con su probabilidad de ocurrencia, amenazas expuestas, así como el impacto negativo que ocasione a las operaciones de negocio. (WCruzy, 2014)
- Seguridad: Es una forma de protección contra los riesgos. (WCruzy, 2014)

ii. Preocupaciones en cuanto a privacidad

Existe una preocupación, que se expande más allá de la comunidad de internet, sobre las consecuencias que un concepto como internet de las cosas trae en cuanto a privacidad. Por ejemplo, la información que recolecta cada uno de los electrodomésticos mencionados arriba, tiene un valor inmenso para análisis de mercados, y trae consigo la inevitable pregunta: ¿para quién más tiene valor esta información? (Castro, 2014)

Por otro lado está la pregunta de qué pasaría si un hacker accede a alguno de estos smart-algo, que quizá no tenga consecuencia en el caso de que acceda los electrodomésticos de la persona promedio, sin embargo está el aspecto de sistemas que controlan aspectos de ciudades enteras. (Castro, 2014)

iii. Beneficios a futuro del internet de las cosas

La idea que subyace bajo el concepto de internet de las cosas (IoT) es muy simple. Y su aplicación, muy difícil. Si todas las latas, libros, zapatos o partes de un vehículo estuvieran equipados con dispositivos de identificación minúsculos, la vida cotidiana en nuestro planeta sufriría una transformación. (Castro, 2014)

Ya no existirían cosas fuera de stock o productos perdidos, porque nosotros sabríamos exactamente lo que se consume en el otro lado del planeta. (Castro, 2014)

El robo sería una cosa del pasado, sabríamos dónde está el producto en todo momento. Lo mismo podría aplicarse a los paquetes perdidos. El internet de las cosas debe codificar de 50 a 100.000 millones de objetos y seguir el movimiento de estos. Para que nos hagamos una idea, una persona puede estar rodeada de 1.000 a 5.000 objetos. (Ecointeligencia, 2013)

Hay un gran potencial en conectar lo que está desconectado ante este desafío que estamos acometiendo, nos gustaría mostrar algunas de las claves sobre internet de las cosas (IoT) y el impacto que va a tener en nuestras vidas, tanto a nivel del diseño sostenible como en nuestro entorno personal. (Ecointeligencia, 2013)

Internet de las cosas (IoT) creará oportunidades valoradas en 14.4 billones de dólares. La combinación de mayores beneficios y bajos costes ayudará a empresas e industrias en el intervalo 2013 de 2022. (Ecointeligencia, 2013)

iv. Principales factores que alimentan a internet de las cosas

- Utilización de activos (reducción de costes) por importe de 2.5 billones de dólares.
- Productividad de los empleados (mayor eficiencia en las tareas) por importe de 2.5 billones de dólares.
- Cadena de aprovisionamiento y logística (eliminación de gastos) por importe de 2.7 billones de dólares.
- Experiencia de usuarios (aumento de clientes) por importe de 3.7 billones de dólares.
- Innovación (reducción del tiempo de llegada al mercado) por importe de 3.0 billones de dólares.

v. Tendencias tecnológicas

Las tendencias tecnológicas son predicciones del nivel de utilización de alguna tecnología, con base en los niveles del consumo, aplicación, factibilidad y utilización que incluyen la nube, la movilidad, Big data, el incremento de capacidad de proceso, y las económicas (como la Ley de Metcalfe) están dirigiendo la economía de IoT. (Eduardo, 2011)

Esas tecnologías y tendencias en los negocios están impulsando a internet de las cosas, creando una oportunidad sin precedentes para conectar lo desconectado: personas, procesos, información y cosas. Actualmente el 99.4 % de los objetos físicos que pueden algún día ser parte de IoT todavía no están conectados. (Eduardo, 2011)

Para obtener el mayor valor de IoT, los líderes empresariales deberían empezar a transformar sus organizaciones basándose en los casos de uso clave que constituyen la mayoría del valor que IoT pone en juego. (Ecointeligencia, 2013)

Estos casos incluyen smartgrid, edificios inteligentes, sanidad y monitorización de pacientes, industrias, educación, flotas comerciales de vehículos, marketing y publicidad, juegos y entretenimiento, entre otros. (Ecointeligencia, 2013)

Las capacidades de seguridad robustas (tanto lógicas como físicas) y las políticas de privacidad son características críticas de la economía de IoT. El valor puesto en juego por IoT está basado cada vez más en la amplia adopción de IoT por compañías del sector privado durante la próxima década. (Ecointeligencia, 2013)

Este crecimiento podría ser inhibido si las capacidades de seguridad de la tecnología no son combinadas con políticas y procesos diseñados para proteger la privacidad tanto de las compañías como de los particulares. (Ecointeligencia, 2013)

Y la siguiente oleada del imparable crecimiento de Internet vendrá de la confluencia entre personas, procesos, información y cosas. Todo ello sin olvidar de la sostenibilidad. (Ecointeligencia, 2013)

A continuación la figura 32 muestra la conectividad de las nuevas tendencias tecnológicas en todas las actividades de la sociedad como lo es en el aspecto laboral, académico científico entre otros:

Figura 32. Tendencia tecnológica

Fuente: Sitio Web AVG Antivirus.

Web semántica

a. ¿Qué es la web semántica?

Es un conjunto de actividades desarrolladas en el seno de World Wide Web Consortium W3C orientados a la creación de tecnologías para publicar datos legibles por aplicaciones informáticas. Se basa en la idea de añadir metadatos semánticos y ontológicos a la world wide web.

Su objetivo es mejorar internet ampliando la interoperabilidad entre los sistemas informáticos usando «agentes inteligentes» , estos son programas en las computadoras que buscan información sin operadores humanos. (Adobe Systems Software, 2014)

b. ¿Para qué sirve la web semántica?

La Web ha cambiado profundamente la forma en la que se comunica, se hacen negocios y realizan su trabajo. Se puede realizar transacciones económicas a través de Internet, logrando el acceso a millones de recursos independientemente de la ubicación geográfica e idioma. (W3C, 2014)

Todos estos factores han contribuido al éxito de la Web. Sin embargo, al mismo tiempo, estos factores que han propiciado el éxito de la Web, también han originado sus principales problemas: sobrecarga de información y heterogeneidad de fuentes de información con el consiguiente problema de interoperabilidad. (World Wide Web Consortium, 2014)

La web semántica ayuda a resolver estos dos importantes problemas permitiendo a los usuarios delegar tareas en software. Gracias a la semántica en la Web, el software es capaz de procesar su contenido, razonar, combinarlo y realizar deducciones lógicas para resolver problemas cotidianos automáticamente. (World Wide Web Consortium, 2014)

c. ¿Cómo funciona la web semántica?

La Web tiene la capacidad de construir una base de conocimiento sobre las preferencias de los usuarios que, a través de una combinación entre su capacidad de conocimiento y la información disponible en Internet, sea capaz de atender de forma exacta las demandas de información por parte de los usuarios en relación, por ejemplo, a reserva de hoteles, vuelos, médicos, libros. (W3C, 2014)

Si esto ocurriese así en la vida real, el usuario en su intento, de encontrar todos los vuelos a Praga para mañana por la mañana obtendría unos resultados exactos sobre su búsqueda. Sin embargo, la realidad es otra. (World Wide Web Consortium, 2014)

Los resultados que se obtendrían con el uso de cualquier buscador actual que ofrecería información variada sobre Praga, pero que no tiene nada que ver con lo que realmente el usuario buscaba. El paso siguiente que el usuario debe realizar es una búsqueda manual, entre esas opciones que aparecen, con la consiguiente dificultad y pérdida de tiempo. (World Wide Web Consortium, 2014)

Con la incorporación de semántica a la Web los resultados de la búsqueda serían exactos, dichos resultados ofrecerían al usuario la información exacta que estaba buscando. La ubicación geográfica desde la que el usuario envía su pregunta es detectada de forma automática sin necesidad de especificar el punto de partida, elementos de la oración como «mañana» adquirirían significado, convirtiéndose en un día concreto calculado en función de un «hoy». (Trapote, 2013)

Algo semejante ocurriría con el segundo «mañana» que sería interpretado como un momento determinado del día. Todo esto a través de una Web en la que los datos pasan a ser información llena de significado. El resultado final sería la obtención de forma rápida y sencilla de todos los vuelos a Praga para mañana por la mañana. (Trapote, 2013)

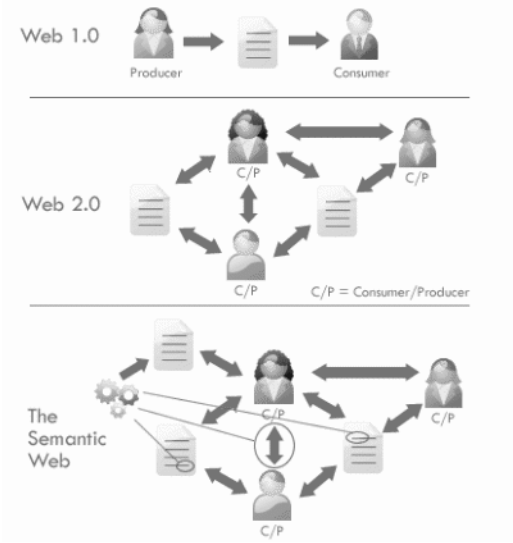


Figura 33. Evolución de la Web.
Fuente: Sitio Web ADR Formación.

CAPÍTULO II. IMPLEMENTACIÓN DE LA INNOVACIÓN

A. Objetivos

Objetivo general

Elaborar un modelo tecnológico en el que se refleje el uso de herramientas Big Data para almacenar, procesar y analizar grandes cantidades de datos para obtener análisis de datos que puedan ayudar en la toma de decisiones de cualquier empresa, independientemente del rubro que manejen.

Objetivo específico

- Identificar los dataset públicos en la nube que contengan una gran cantidad de registros para utilizarlos en el procesamiento y almacenamiento de la información.
- Proponer herramientas tecnológicas que permitan el procesamiento y análisis de los datos como Hadoop con su herramienta Hive y R.
- Analizar y presentar los resultados obtenidos utilizando herramientas de visualización como Google Chart, Jqplot y D3.js

- Diseñar la propuesta de un modelo tecnológico para la pequeña y mediana empresa (pymes) aplicando herramientas de Big Data para el análisis de datos.

B. Diseño de la innovación

En El Salvador se desconoce información sobre el Big Data, pero existe mucho interés de incursionar y utilizar las herramientas que faciliten el procesamiento y análisis de grandes volúmenes de datos.

En este proyecto se decidió hacer uso de metodologías que permitan describir los pasos necesarios para utilizar algunas herramientas de Big Data y los requisitos que deben considerarse para poder hacer uso de ellas.

En términos generales la metodología consistió en lo siguiente:

1. Almacenar y procesar un dataset público haciendo uso de Hadoop, el cual contiene información sobre registro de empresas de servicios pertenecientes a las Mypes.

2. Con la herramienta Hive que viene en la distribución Hortonworks de Hadoop se harán las consultas necesarias, pues esta herramienta es similar a las instrucciones que se utilizan en SQL, para los que están acostumbrados a trabajar con base de datos relacionales les será fácil entender la lógica de cómo trabaja Hive, y en El Salvador SQL es el software más utilizado para bases de datos y esa es la razón por la que se seleccionó esta herramienta.

3. Para el análisis de datos estadístico se utilizará el programa R, haciendo uso de los dataset de servicios. Las razones del porque se seleccionó este programa se debe a que devolverá información sobre datos importantes de los productos que están almacenados y, a la vez, permitirá que se realicen conclusiones con base en los resultados.

4. Después de haber hecho un análisis estadístico y las consultas pertinentes de los datos se procedió a realizar los gráficos necesarios para una mejor comprensión de los resultados y obtener conclusiones y tomar decisiones, las herramientas utilizadas son Google Chart, Jqplot y D3.js.

Para la implementación de la metodología propuesta fue necesario tomar en cuenta ciertos requisitos técnicos relacionados con el hardware.

C. Metodología y estrategia

En este capítulo se trató de explicar los procesos fundamentales que fueron necesarios implementar para lograr el objetivo planteado. Estos deben ser de forma ordenada iniciando desde la elección del dataset público y luego con el uso de las herramientas necesarias para el almacenamiento, procesamiento, análisis y visualización de los datos.

Hadoop es la única herramienta que requiere de muchos recursos de hardware, sobre todo de memoria RAM y de disco duro, y por ello lo más recomendable es que esté instalado en un servidor; pero por las limitantes encontradas y no disponer de suficientes recursos en el equipo utilizado, se optó por crear máquinas virtuales para simular un master y un esclavo y hacer el despliegue en un entorno similar a uno de producción. (Cloud Google Inc.2016)

Los demás programas utilizados son gratuitos y con lecturas de manuales en la web, se pueden llegar a utilizar sin ningún problema.

D. Requisitos técnicos

Herramientas que faciliten el análisis de ellos.
Los requisitos técnicos y las tecnologías utilizadas son las siguientes:

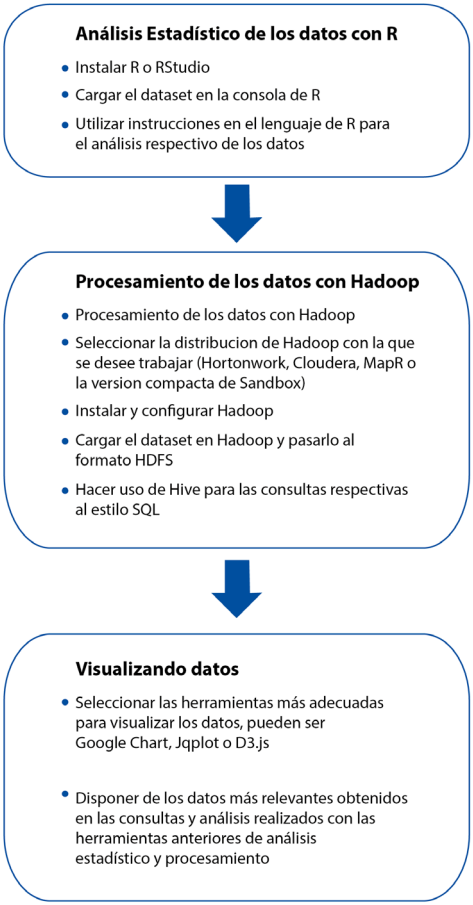
Tabla 3: Requisitos técnicos

Herramienta tecnológica	Requisito de hardware	Requisito de software
Hadoop	• 5.ª generación del procesador Intel® Core™ i7 • Memoria de 16 GB expandible a 32 GB • Disco duro de 1 a 2 TB	• Windows o Linux • Linux de 64 bits • Oracle VM VirtualBox • CentOS 6.5 • Distribución • Hortonworks
Programa R	No requiere de características especiales.	• Windows, Mac o Linux • R o R Studio

Herramienta tecnológica	Requisito de hardware	Requisito de software
Google chart, Jqplot, D3.js	No requiere de características especiales.	• Cualquier sistema operativo • Cualquier editor de texto: bloc de notas, sublime text o notepad ++ • Navegador web • Servidor web

E. Modelo del proceso

El siguiente esquema muestra los procesos operativos del análisis y visualización de datos para comprender el modelo de Big Data.



Para este caso, como se ha mencionado en capítulos anteriores, se hizo uso de dos dataset que contienen registros de empresas de servicios. Estos ficheros serán manipulados de tal manera que pueda ser almacenados, uno de ellos con Hadoop y luego con la herramienta Hive poder hacer las consultas necesarias y otros dataset será utilizado en el programa R, se hará análisis estadísticos de los datos más relevantes del fichero seleccionado.

Con la información obtenida se procedió a elaborar los gráficos, haciendo uso de cualquiera de las herramientas de visualización mencionadas anteriormente que reflejarán lo más importante de los datos para obtener conclusiones que servirán en la toma de decisiones.

Ficheros utilizados

Los ficheros están en formato csv, los cuales han sido descargados de la web.

Almacenando y procesando datos con Hadoop

Como se dijo antes, Hadoop es un proyecto de software libre con licencia Apache, cuya finalidad es prestar una plataforma para la gestión de grandes cantidades de datos. Los principales componentes que constituyen Hadoop son el sistema de archivos HDFS y el motor MapReduce.

Por lo tanto, se procedió a instalar CentOS que es la opción recomendada para desplegar posteriormente Hortonworks y para este proyecto se hará uso de la herramienta Hive para realizar las consultas respectivas de uno de los dataset, el cual tiene registro de medicamentos, podrían ser las siguientes:

- Mostrar el principio activo del medicamento, nombre del producto y el laboratorio que lo fabrica.
- Mostrar los medicamentos y la fecha de resolución.
- Mostrar los medicamento y para que tratamiento son usados.
- Mostrar medicamentos que son usados para cualquier tipo de enfermedad que interese consultar.
- Mostrar los medicamentos con sus números de registro.
- Mostrar la cantidad de medicinas que hay para un tratamiento específico.
- Mostrar la cantidad de titulares o fabricantes de los medicamentos.

Analizando datos con R

Dentro de los objetivos se encuentra la utilización del programa R para el análisis estadístico de los datos; por lo tanto, hay que importar el dataset al programa R y comenzar a realizar los análisis estadísticos respectivos.

Visualizando datos

Cada día se generan millones de datos e incluso de forma individual estamos inundados de correos electrónicos, fotografías videos, música, libros electrónicos y aparte de ello se almacena más información y documentos en la nube, por ejemplo Google Drive o mediante cualquier disco virtual, lo cual hace que exista demasiada información y de la cual no se es capaz de conocer el poder que esta puede tener.

Así mismo, si se añade el concepto de *Big Data*, «grandes conjuntos de datos (o datasets)», se podrá comprender que la principal dificultad no es la captura y almacenamiento de los datos, sino el análisis y su posterior representación visual estática o interactiva.

Los datos por si solos no tienen sentido, a menos que estos se conviertan en información útil y comprensible para luego acceder al conocimiento.

Se sabe que los datos son la materia prima de la información y la información la materia prima del conocimiento y, por supuesto, al tener conocimiento se pueden tomar decisiones.

Para las empresas el tiempo en procesar todos los datos generados es valioso, pero se sabe que no se puede lograr sin las herramientas adecuadas que permitan transformar la multitud de datos. Una vez se haya logrado transformar los datos en información, es necesario presentarlos de manera atractiva y que sea fácil de comprender, por lo tanto, la magia de la visualización de información radica en la captura y síntesis previa de la información y mediante el uso de diversas técnicas visuales como diagramas, gráficas, esquemas, nubes palabras, conexiones, grafos, pueda ser transformada y con ello facilitar la compresión de la misma.

Es por ello, después de haber hecho uso de herramientas para el almacenamiento, procesamiento y análisis de los datos, es necesario hacer uso de herramientas de visualización que permitan comprender los resultados; pero en un formato gráfico, debido a que para la mente humana es mucho más fácil entender imágenes que únicamente texto o números.

De esto se encarga la visualización de datos o conocido como el diseño de la comprensión. Los datos deben ser comprendidos de manera efectiva. El objetivo de toda buena visualización tiene que centrar la atención del interesado de la información en aquello que realmente es relevante e importante.

Una vez se disponga de ese nuevo conocimiento, este permitirá hacer análisis y sacar conclusiones que serán importantes para la toma de decisiones como se observa en la figura 34.

Figura 34. Diagrama de estructura de los datos hacia la sabiduría.

Fuente: Propia.

Lo que se pretende con el diagrama anterior es pasar los datos a la sabiduría, porque de eso dependerá tomar buenas decisiones.

Las herramientas de visualización que se utilizaron para este proyecto fueron las siguientes: Google chart, Jqplot o D3, cada una de ellas tienen sus propias características, como se mencionó en capítulos anteriores, por lo que dependerá del uso que quiera darse en la representación gráfica de los datos, así será la selección de cualquiera de ellas o se pueden utilizar las tres si se desea.

F. Recursos y presupuesto

Se contó con el recurso humano idóneo; tres profesionales expertos en el área de sistemas informático y un grupo de estudiantes de la facultad de Ingeniarías. En el equipo de investigadores se contó con un especialista en el área de Big Data. También, se cuenta con el apoyo de la Vicerrectoría de Investigación

Cronograma de actividades

Se contó con un equipo de tres ingenieros en sistemas computacionales que han distribuido el tiempo por medio un cronograma de actividades que se desarrollan desde enero a noviembre de 2016.

CAPÍTULO III. RESULTADOS DE LA INNOVACIÓN

A. Cambios en necesidades y problemas abordados

Debido a la gran notoriedad que está teniendo esta tendencia y parte de las nuevas tecnología. Para cualquier persona sin o con conocimientos tecnológicos, se pregunta cómo se almacena toda la información que se genera en el mundo: como son las redes sociales Facebook, Twitter, Smartcities, Istagram o como Google es capaz de manejar todas las transacciones que se hacen a diario. Pero no solo se queda aquí, pues el Big Data alcanza todos los ámbitos: bolsa, climatología, astronomía, marketing, entre otros. Por lo que la cantidad de datos que se genera actualmente es abrumadora y solo el hecho de saber cómo se consigue captar y analizar dicha información, parece una justificación bastante razonable para buscar herramientas que proporcionen soluciones atractivas.

Por otra parte el almacenamiento de la información cada día se incrementa por esta razón se ha decidido implementar nuevas tecnologías que cumplan con los requisitos de las grandes empresas, pues almacenan cantidades enormes de información y requieren de mecanismos que les permitan realizar sus procesos de forma rápida y eficiente.

En la actualidad, la tecnología del Big Data está tomando cada vez más realce dentro del mundo de los negocios y las estrategias, el conocimiento de esta tecnología puede ser aprovechada por cualquier empresa para ofrecer una mejor forma de brindar sus productos y servicios.

B. Cambios observados en el bien, servicio o proceso que se innovó

El aprovechamiento de tecnología del Big Data permite que las empresas conozcan más de cerca a sus clientes, prestarle un mejor servicio, mejorar la calidad de sus productos, generar oportunidad para ingresar a nuevos mercados, completar su portafolio de clientes, entre otras tareas que generen beneficios al negocio.

Por lo tanto, la investigación sobre la tecnología de Big Data y el uso de herramientas que faciliten el procesamiento, análisis y visualización de los datos, está basada en los siguientes indicadores: explorar los conocimientos que se tienen sobre Big Data, uso de las distintas herramientas para el procesamiento de los datos, conocimiento de herramientas de visualización de grandes cantidades de datos, conocer las preferencias y los elementos necesarios que se pueden utilizar para mejorar los procesos en las empresas.

Por lo tanto, este trabajo de investigación, deja un proceso metodológico para que las empresas pymes puedan acceder al uso de herramientas tecnológicas Big Data para facilitar el almacenamiento, análisis y visualización de grandes volúmenes de información para la toma de decisión.

C. Pruebas y demostraciones de la eficacia, eficiencia y efectividad del proyecto de innovación

En esta etapa del proyecto se realizaron las pruebas de análisis de los dataset seleccionado con el fin de desarrollar un plan de implementación para una mediana y pequeña empresa.

Los diferentes elementos del modelo se podrán realizar aproximadamente de la siguiente forma **(Ver Tabla)**.

Tabla 4: Requerimientos técnicos

Número	Actividad	Duración	Descripción
1	Definir el objetivo de la implementación del modelo	3 días	Determinar el impacto esperado con la implementación del modelo en la organización.
2	Definir equipo ejecutivo	1 día	Determinar a los miembros del equipo directivo del proyecto, estos deben de ser personal con capacidad de gestión y compenetrados el cumplimiento del objetivo.

3	Diagnóstico de necesidades de información transversal en toda la organización	3 días	Estudio resultado del análisis de la misión y visión de la empresa y su relación con el objetivo del proyecto.
4	Definir al equipo de apoyo interfuncional	3 días	Identificar a los miembros claves del proyecto para fluir la información interfuncional.
5	Definición de estrategias de comercio	5 días	Se buscó definir la nueva estrategia de mercadeos a partir de la información proporcionada por el análisis de los clientes, en ese sentido, se buscara que el departamento de mercadeo ofrezca alternativas de operativizar el comercio con los clientes.
6	Determinación de necesidades (tecnológicas, legales y organizacionales de la empresa)	5 días	Equipo de TI y equipo legal, analizaran si existe alguna legislatura interna que contradiga los resultados esperados.
7	Capacitación del personal técnico	10 días	En esta actividad se pretende dar a conocer a el personal de tecnología la forma de operar en el uso de la tecnología del BigData y la forma aprovechar los diversos recursos que esta ofrece.

8	Actualización de datos de clientes	10 días	Se debía de actualizar la información de las base de datos relacionadas a los clientes a los que nos permitirá aumentar y retener para hacer el comercio electrónico proporcionándoles valor agregado a la oferta. Este factor es muy importante realizarlo y se pide centralizar esfuerzos en cumplir dicho requerimiento. La información que se actualice será el punto de entrada para el análisis de las preferencias del cliente así como su historial en la web.
9	Adquisición de hardware y software especializado	15 días	Proceso de compra del hardware y software para operar la tecnología de BigData. Si se posee recursos, se buscara actualizarlo para un mejor rendimiento. Hadoop es la tecnología propuesta.

10	Instalación y configuración del software especializado	5 días	El software especializado debe de configurarse para trabajar de forma integrado, para esta actividad se requiere que se trabaje con una contra parte especializada que permita hacer la comparación de los resultados obtenidos.
11	Ajustes de sistemas y capa de seguridad	10 días	Los Sistemas transaccionales deben de ser actualizados por los miembros del departamento de TI a través de una capa de seguridad, esta debe de ser la misma que se aplique a todos con el objetivo de perdida de la información. No es una requerimiento indispensable pero si es válido que se coloque.
12	Ajuste de sistemas transaccionales de gestión de clientes	15 días	Cuando los datos han sido actualizados, los nuevos Sistemas deben de ser capaces de poder interactuar con la nueva información. Esta actividad puede obviarse si se actualizaron junto con los datos de los clientes. Elaboración de un pequeño portal de consultas.

13	Monitoreo de tráfico de datos de clientes	5 días	Proyección de volumen de crecimiento de los datos de los clientes para interferir en el almacenamiento de los mismos a partir de diversos orígenes de datos.
14	Extracción de información de datos de trafico de clientes	5 días	Explorer los algoritmos de búsqueda referencial. Tarea realizada por el departamento de TI para buscar patrones en base a las reglas definidas.
15	Generación de informes de análisis de datos	10 días	Formatear los primeros reportes del escaneo de clientes a partir de patrones definidos.
16	Capacitación de personal usuario	3 días	Uso de las diferentes unidades involucradas de la organización para el uso de los datos a través de un pequeño portal diseñado para tal fin.
17	Pruebas de resultados	5 días	Análisis de resultados obtenidos a partir del portal.
18	Puesta en marcha	3 días	Preparación de la puesta en marcha.

Fuente: Propia.

Con un tiempo aproximado de 4 meses (mayo-agosto 2016) algunas actividades se realizaron en paralelo y otras fueron opcionales dependiendo del grado de avance en función de los requerimientos definidos. De igual manera el presupuesto

Validación del modelo

La propuesta que se ha diseñado se demostró con un análisis mediantes pruebas de eficacia; se consideraron los aspectos relacionados a recursos humanos, aspectos legales y tecnológicos mencionados en el modelo y la disponibilidad de estos en las empresas para poder determinar con estas la idoneidad de implementar el modelo.

Los principales obstáculos que se encontraron para poder validar el modelo fue que las empresas no cuentan a corto plazo con estrategias definidas para el aprovechamiento del Big Data para generar comercio electrónico, tampoco cuenta con los recursos a la fecha de los diferente componentes propuestos en el modelo; sin embargo, se mostró un gran interés en implementar el modelo a mediano plazo.

Este documento se diseñó como guía para las empresas que puedan identificar todos aquellos aspectos que son el fundamento para incorporar el comercio electrónico como una herramienta base en sus negocios.

D. Percepciones y evaluaciones de usuarios y beneficiados

Con base en los resultados obtenidos de la investigación se alcanzó a las pequeñas y medianas empresa para darle a conocer los beneficios que se pueden obtener tomando en cuenta el uso de las herramientas para Big Data.

Se espera alcanzar a un grupo de empresa que quieran implementar el uso de esta tecnología para obtener mayor eficiencia en la toma de decisiones. A continuación se detalla la lista de beneficiarios:

Beneficiarios

- La administración de las empresas pequeñas y mediana
- Los estudiantes y docentes de la facultad de Ingenierías de la UEES
- Posicionamiento para la Universidad Evangélica de El Salvador en el área de uso de herramienta Big Data

CAPÍTULO IV. CONCLUSIONES Y RECOMENDACIONES

A. Conclusiones

- Se elaboró un procedimiento tecnológico que permitiera reflejar un buen uso de las herramientas Big Data, que son novedosas y no se conocen las herramientas que se utilizan para el almacenamiento, procesamiento y análisis de grandes volúmenes de datos y es por ello que con las metodologías propuestas se ha tratado de dar a conocer el funcionamiento de alguna de ellas, para motivar a las empresas a involucrarse en las nuevas tecnologías.
- Se identificaron los dataset públicos en la nube que contenían gran cantidad de registros permitiéndonos procesar y analizar la información almacenada.
- Se mostraron herramientas para dar uso de Hadoop para el almacenamiento y procesamiento de los datos para luego poder hacer uso de la herramienta Hive y por medio de ella realizar consultas al estilo SQL. Esta herramienta involucra implícitamente el proceso Map Reduce, lo cual facilita aún más su uso.
- Se analizaron y se presentaron los resultados obtenidos con las herramientas de procesamiento y análisis, se ha hecho uso de tres de ellas: Google Chart, Jqplot y D3.js, por supuesto que no es necesario trabajar con todas, sino que dependerá de las necesidades a representar y de las ventajas que ofrece cada herramienta, así como de la habilidad que se tenga en la programación con Java Script, pero como se ha mencionado a la hora de trabajar con las herramientas de visualización, existen ejemplos y galerías en la web sobre cada una de ellas que se pueden reutilizar y adaptarlas a las necesidades de cada empresa u organización.
- Se diseñó la propuesta de un modelo apropiado para las pymes, aplicando las herramientas Big Data para el análisis de datos.

B. Recomendaciones

El uso de la información proveniente de la pymes analizadas con herramientas de Big Data y aplicando el modelo propuesto, le permitirá a la empresa conocer patrones conductuales de los potenciales clientes que le faciliten a las pymes

llegar a tomar decisiones efectivas, respecto a los productos y servicios que le puedan llegar a ofrecer a sus clientes según sus necesidades.

Diagnosticar las necesidades de las pymes para obtener los datos idóneos para el análisis de los dataset.

Implementar herramientas actualizadas de administración de datos Big Data.

Aplicar correctamente las herramientas de visualización de los datos.

Se recomienda aplicar la propuesta tecnológica en una pymes para utilizar la información de su Base de Datos para lo cual es necesario establecer un convenio entre la pymes y la UEES para garantizar la confidencialidad de la información.

Una vez teniendo el convenio con la pymes se podría tener la base necesaria para una siguiente investigación en la cual se podría aplicar los sistemas de Seguridad de Big Data.

C. Plan de socialización de resultados

- Presentación de comunicaciones orales, poster científico y conferencias magistrales.
- Presentaciones a estudiantes de la Facultad de Ingenierías de la UEES. (En la bienvenida de Ciclo I-2017).
- Autoridades de la administración pymes (Invitación a representantes e interesados).
- Presentación en certamen de Investigación de cátedra de la Facultad de Ingenierías 2017.
- Presentación en congreso internacional de la Vicerrectoría de Investigación UEES 2017.
- Ponencias en instituciones de educación superior.
- Publicaciones: artículo que sea la base para la elaboración de un libro.

FUENTES DE INFORMACIÓN CONSULTADAS

1. Adobe Systems Software. (2014). Predictive Intelligence. Obtenido de: https://www.adobe.com/es/marketing-cloud/web-analytics/premium-predictive-intelligence.html
2. Aguilar, L. J. (2013). Big Data, Analisis de los grandes volumenes de datos. Mexico: Alfaomega.Adobe. (2014). Web Semantica. Mexico: N/A. Obtenido de http://es.wikipedia.org/wiki/Web_sem%C3%A1ntica
3. Analytics Software & Solutions. (2014). Obtenido de: SAS® Analytics https://www.sas.com/en_us/software/analytics.html
4. Bernardo, A. (8 de mayo de 2013). Think Big. Obtenido de http://blogthinkbig.com/big-data-cancer/
5. Bustillo, I. (2013). Business Intelligence Obtenido de Pentaho: http://ignaciobustillo.focalrock.com/blog/blog/63-ique-es-pentaho
6. Carrillo Ruiz, J. A., Marco de Lucas, J. E., Dueñas López, J. C., Cases Vega, F., Fernández, J. C., Pereda Laredo, L. F., & González Muñoz de Morales, G. (marzo de 2013). BIG DATA en los entornos de Defensa y Seguridadhttp://www.ieee.es/. Obtenido de http://www.ieee.es/Galerias/fichero/docs_investig/DIEEEINV03-2013_Big_Data_Entornos_DefensaSeguridad_CarrilloRuiz.pdf
7. Castro, L. (2016 de 3 de 2014). Aprender Internet. Obtenido de: http://aprenderinternet.about.com/od/ConceptosBasico/a/Internet-de-las-cosas.htm
8. Centro de Comercio Internacional UNCTAD/OMC,(2002) Clave del Comercio Electrónico: Guía para Pequeños y Medianos Exportadores. Colombia: CCI. XI, 291 págs.
9. Cloud Google Inc. (2016). Tres ejemplos reales de uso inteligente del Big Data Obtenido de: http://www.muycomputerpro.com/2015/02/24/ejemplos-reales-uso-inteligente-big-data
10. Cloudera, Inc. (2016). Foro Acerca de Cloudera Obtenido de: http://es.cloudera.com/about-cloudera.html
11. De Juana, R. (24 de febrero de 2015). MuyComputerPro.

12. Dean, MapReduce: Simplified Data Processing on Large, (2004). Obtenido de:http://static.googleusercontent.com/media/research.google.com/es//archive/mapred

13. Ecointeligencia. (23 de 10 de 2013). Claves del Internet de las Cosas.

14. FondosFidelity. (2012). Big data: una "revolución industrial" en la gestión de los datos digitales. Obtenido de: https://docs.google.com/file/d/0B8alfTMTeme2ckVmTEw5bHdYOGc/view

15. Gerencia, v. (13 de 03 de 2014). Funciones gerenciales. Recuperado el 2 de 2014, de http://www.degerencia.com/tema/comercio_electronico

16. Ghemawat, S., Gobioff, H., & Leung, S.-T. (octubre de 2003). The Google File System. In Proceedings of the 19th ACM Symposium on Operating Systems Principles. Obtenido de http://static.googleusercontent.com/media/research.google.com/es//archive/gfs-sosp2003.pdf

17. Google Chart Inc. (2016).Dynamic Data. Obtenido de: https://developers.google.com/chart/interactive/docs/queries

18. Hortonworks Inc. (2016). Big Data Analytics At A Tipping Point. Obtenido de: http://es.hortonworks.com/blog/big-data-analytics-tipping-point/

19. http://aplicaciones.digestyc.gob.sv/Clasificadores/Sistema/Documentos/DocumentoCLAEES.pdf

20. http://es.scribd.com/doc/54550614/Como-Funciona-La-Web-Semantica

21. http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

22. http://www.sas.com/offices/latinamerica/mexico/technologies/analytics/

23. International Business Machines Corp. (12 de 10 de 2013). Herramientas framework para Big Data. Obtenido de: http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

24. International Business Machines Corp. (18 de 6 de 2012). Componente de una plataforma Big Data. Recuperado el 2014, de http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

25. International Business Machines Corp. (18 de 6 de 2012). Tipos de datos de Big Data, Web y redes sociales. Recuperado el 2014, de http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

26. Jedox. (2013). Business Intelligence - innovadora, simple y móvil.

27. Joyanes Aguilar, L. (2013). Big Data: Análisis de grandes volúmenes de datos en las organizaciones. Mexico: AlfaOmega.

28. Lozano, J. F. (18 de 2 de 2014). Ingenieria de software. Obtenido de HIVE - Consultas “tipo SQL” sobre Hadoop: http://www.franciscojavierpulido.com/2013/11/hive-consultas-tipo-sql-sobrehadoop.html

29. Lozoya, J. (15 de Julio de 2013). Clase de comercio electronico, B2E. Recuperado el 2014, de Comercio electronico: http://suite101.net/article/clases-de-comercio-electronico-b2b-b2c-b2ab2e-c2c-c2g-b2g-a26589

30. MapR Technologies, Inc. (2016). The MapR Advantage. Obtenido de: https://www.mapr.com/why-hadoop/why-mapr

31. Martínez, J. E. (2013). Desafío y Oportunidades de las Pymes Salvadoreñas. San Salvador: N/A.

32. Ministerio de Economía .MINEC. (2012). Estadísticas y Censos El Salvador. San Salvador. Obtenido de:

33. Obtenido de: http://tendenciasytecnologiasdeti.blogspot.com/2011/10/tendencias-tecnologicas.html

34. Obtenido de Big Data: http://ramotecno.blogspot.com/

35. Obtenido de http://www.ecointeligencia.com/2013/09/6-claves-internet-de-las-cosas-iot/

36. Obtenido de Palo Suite: http://translate.google.com.sv/translate?hl=es&sl=en&u=http://www.palo.net/&prev=/search%3Fq%3Dque%-2Bes%2BPalo%2BBI%2BSuite/Jedox%26es_sm%3D122%26biw%3D1366%-26bi h%3D643

37. Obtenido de: http://www.muycomputerpro.com/2015/02/24/ejemplos-reales-uso-inteligente-big-data

38. Olguín, F. P. (30 de Junio de 2013). Tecnología y Computación Informática.

39. Oracle. (15 de 8 de 2013). Base de datos orientada a grafos. Obtenido de http://www.javamexico.org/blogs/ezamudio/neo4j_base_de_datos_orientada_grafos

40. Pontigo, J. S. (30 de Julio de 2013). Tecnología y Computación Informática. Obtenido de Big Data: http://ramotecno.blogspot.com/

41. Ramírez, E. (12 de 10 de 2011). Tendencias y Tecnología.

42. Rosa, V. I. , Rivera, J. G. ,(2016) Big Data, Análisis de Datos en La Nube. El Salvador: Universidad Tecnológica de El Salvador.

43. Scherer, M. (9 de noviembre de 2012). DataPrix. Obtenido de http://www.dataprix.com/noticias-it/tendencias-tecnologicas/big-data/big-data-ayudaron-obama-ganar-las-elecciones

44. Souto, S. (4 de mayo de 2015). Hechos de Hoy. Obtenido de http://www.hechosdehoy.com/big-data-pilar-fundamental-para-el-desarrollo-de-las-ciudades-inteligentes-43285.htm

45. tStratebi Inc. (2014). Obtenido de Talend Open data solutions: http://www.stratebi.com/talend

46. TestingSoft. (2014). Tipos de datos a explorar. Recuperado el 2014, de http://www.evaluandosoftware.com/nota-3684-Que-es-el-Big-Data.html.

47. The Apache Software Foundation (2014). Welcome to Apache Obtenido de: http://hadoop.apache.org/

48. The R Foundation (2016). The R Project for Statistical Computing Obtenido de: https://www.r-project.org

49. Trapote, P. (2013).Como Funciona La Web Semantica. Obtenido de: https://www.r-project.org

50. WCruzy. (20 de 2 de 2014). Seguridad de la Informacion. Obtenido de http://wcruzy.uphero.com/iii/seguridadinformacion.pdf

51. World Wide Web Consortium. W3C.15 de 3 de 2014).Guias Breves. Obtenido de http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica

52. Zikopoulos, P. (2015). BeSmart. Obtenido de http://www.besmart.company/blog/big-data-transforma-ciudades/

53. Zoho Corp. (2016). Advanced Analytics for IT. Obtenido de: https://www.manageengine.com/it-analytics.html