

Análisis de la viabilidad de la implementación de redes Big Data en Colombia

Universidad Distrital FJDC Facultad Tecnológica

Presentado por:

**Jorge Eduardo Reíta Reyes
Héctor Javier Salinas Hernández**

Director de Monografía

ING. JOSE DAVID CELY CALLEJAS

Ingeniería en Telecomunicaciones Bogotá, Agosto 11 de 2016

TABLA DE CONTENIDOS

Contenido

TABLA DE CONTENIDOS.....	2
CAPITULO 1	10
ESTADO DEL ARTE.....	10
1.1 BIG DATA EN COLOMBIA	10
1.2 PROPUESTAS DE BIG DATA EN EL MINTIC	10
1.3 TENDENCIAS DE BIG DATA EN COLOMBIA PARA PYME	11
1.3.1. BIG DATA EN ALMACENES, FIRMA SENSETA:	11
1.3.2. SENSETA TERRAINLOGICS:	11
1.3.3. TODOESTADISTICAS COLOMBIA:	12
1.3.4. SAS EN COLOMBIA:	12
1.3.5. DATA CENTER DE IBM EN COLOMBIA:	12
CAPITULO 2	13
MARCO TEORICO.....	13
2.1 ¿QUE ES BIG DATA?.....	13
2.2 FUENTES DE DATOS:	14
2.3 TIPOS DE DATOS	15
2.3.1. REDES SOCIALES:.....	16
2.3.2. MÁQUINA A MÁQUINA (M2M):	16
2.3.3.- TRANSACCIONES:	16
2.3.4. BIOMÉTRICA:.....	16
2.3.5. GENERADAS POR LOS HUMANOS:.....	17
2.4 LA IMPLEMENTACIÓN REVOLUCIONARIA DE BIG DATA	17
2.5. LA ALTERNATIVA HÍBRIDA A LA IMPLANTACIÓN DE BIG DATA	18
2.6 PANORAMA MUNDIAL	19
2.6.1. ÁMBITO INTERNACIONAL	19
2.6.2. ÁMBITO EUROPEO	21
2.6.3. ÁMBITO ASIÁTICO.....	22
2.6.4. ÁMBITO AMERICANO.....	23
2.7 ANÁLISIS ÁMBITO NACIONAL	25
CAPITULO 3	28
COMPONENTES DE UNA PLATAFORMA BIG DATA.....	28

3.1. HADOOP	28
3.2. FUNCIONAMIENTO HADOOP	29
3.3. MAPREDUCE	29
CAPITULO 4	31
BIG DATA COLOMBIA.....	31
4.1 ALIANZA CAOBA (Centro de alianza y aplicación en Big Data y Data Analytics): ..	31
4.2 BBVA Y BIG DATA:	32
4.3 DAVIVIENDA:	32
4.4 BIG DATA Y EL SECTOR DE LA SALUD: EL FUTURO DE LA SANIDAD	33
4.5 EL ESTADO DE LA CUESTIÓN	33
4.6 AREA DE LA SALUD - CONVERSATORIO SOBRE BIG DATA Y GESTIÓN DE LA SALUD EN LAS CIUDADES.....	34
4.7 ALIANZA CAOBA	34
CAPITULO 5	35
VIABILIDAD DE BIG DATA EN COLOMBIA.....	35
5.1 SECTOR AGRÍCOLA.....	35
5.2 SECTOR INDUSTRIAL.....	37
5.3 SECTOR DE COMERCIO Y/O PUBLICITARIO.....	40
5.3.1 INTERNET	42
5.4 SECTOR FINANCIERO	43
5.4.1 ASPECTOS PROMETEDORES EN EL SECTOR FINANCIERO:	46
CAPITULO 6	47
ESTUDIO COMPARATIVO DE TECNOLOGIAS OFRECIDAS EN EL MERCADO PARA BIG DATA EN COLOMBIA	47
CAPITULO 7	57
MODELO DE IMPLEMENTACIÓN DE PROYECTOS DE BIG DATA ORIENTADO A PYME DE COMUNICACIONES	57
7.1 MODELO DESCRIPTIVO PARA ESTABLECER LA VIABILIDAD DE UN PROYECTO BIG DATA	57
7.2 MODELO PROPUESTO PARA IMPLEMENTACIÓN DE PROYECTOS DE BIG DATA ORIENTADO A PYME DE COMUNICACIONES.....	61
7.2.1 Indexación universal en tiempo real de cualesquiera datos de máquina	62
7.2.2 Búsqueda libre y análisis de historial en tiempo real.....	62
7.2.3 Descubrimiento de conocimiento automatizado de los datos	63
7.2.4 Monitoreo de datos y alertas en tiempo real.....	63
7.2.5 Debe proporcionar informes y análisis ad hoc.....	64
7.2.6 La empresa debe tener la capacidad de crear rápidamente paneles de	

instrumentos y visualizaciones personalizadas	64
7.2.7 La empresa deberá usar hardware de producto en la escalabilidad de datos	64
..... ¡Error! Marcador no definido.	
7.2.8 Seguridad y controles de acceso	64
7.2.9 Se debe dar soporte a la implementación flexible y de multi-inquilinos	65
7.2.10 Integración extensible vía APIs	65
CAPITULO 8	66
COSTOS ESTIMADOS PARA PYMES DE COMUNICACIONES	66
8.1 CONSIDERACIONES PARA IMPLEMENTAR BIG DATA (COSTOS)	66
8.1.1 CULTURA ORGANIZACIONAL	66
8.1.2 INTEGRACIÓN DE LAS ÁREAS DE LA ENTIDAD	66
8.1.3 INFRA ESTRUCTURA Y NODOS	67
8.2 COSTOS ESTIMADOS PARA PYMES DE COMUNICACIONES	67
8.3 PRELIMINARES	69
8.4 TABLEROS ACOMETIDAS Y CABLEADO ESTRUCTURADO	71
8.5 ILUMINACIÓN Y TOMAS	72
8.6 EQUIPOS DE CÓMPUTO Y SERVIDOR	72
8.7 CAPACITACIÓN Y CIENTÍFICOS DE DATOS	73
8.8 PLANOS Y CERTIFICACIONES	73
CAPITULO 9	75
CONCLUSIONES	75
CAPITULO 10	77
BIBLIOGRAFÍA	78
<i>AGRADECIMIENTOS</i>	79

INTRODUCCIÓN

Para poder comprender las implicaciones, beneficios y recursos necesarios para la implementación de Big Data en Colombia debemos empezar por esclarecer el significado del concepto inicialmente, sin embargo también es necesario establecer la utilidad del mismo, el por qué se hace necesario implementar este tipo de tecnología en sistemas que han afrontado el manejo de información de una forma relativamente complaciente para sus usuarios hasta la actualidad.

De una u otra forma el avance de la tecnología moderna implica nuevos cambios diariamente en el tipo y cantidad de información que se maneja, el hecho de que la información no se pueda manipular de una manera complaciente y se pueda interpretar como un conjunto voluble y cambiante de datos en el cual es complejo analizar todas las variables y componentes que conllevan, hacen imprescindible que las organizaciones, entidades, instituciones y demás usuarios que manejan niveles considerables de información deban reestructurar sus métodos para el manejo y análisis de información con el propósito de optimizar las operaciones en el procesamiento de los datos para generar respuestas rápidas y precisas y de este modo tener la posibilidad de ofrecer un mejor servicio.

Las operaciones y procesos más eficientes en un índice notable de desarrollo son un resultado que en cualquier entidad se traduce como beneficios económicos lo cual justifica mejorar la calidad en el análisis de la información como se manifestaba con antelación es producto no solo de la expansión de las organizaciones por abarcar mayor parte del mercado sino que también resulta ser una implicación directa del crecimiento de las tecnologías difundidas en la red que van desde aplicaciones de geo-referenciamiento hasta redes sociales o streaming a través de las cuales se ha buscado potenciar el marketing y han sido propulsores en el desarrollo de las empresas. Es por eso que se hace necesario implementar un medio más eficiente para la medición y análisis de la información recolectada desde todo tipo de fuentes y de la cual no se obtiene el mayor beneficio por la ineficiencia de los sistemas para exponerla de manera ordenada, detallada y explícita, inconveniente que se pretende aminorar en gran medida con la implementación de Big Data.

Para el contexto nacional, Big data no es una tecnología que no haya sido implementada anteriormente, no obstante muchas de las entidades referentes en el tema son extranjeras o son sucursales de potencias a nivel mundial que se adentran en el mercado colombiano con el fin de obtener un negocio rentable dado que la información es un recurso que abunda en todo el globo, es por eso que entidades como la MINTIC o Bancolombia apuntan al desarrollo de esta tendencia con el fin de sacar

provecho de todos los beneficios que esta otorga. A lo largo del presente escrito se estudiarán las tendencias mundiales de Big data, el ámbito nacional, la estructura de Big data y otros factores con el propósito de establecer un modelo a seguir para evidenciar la viabilidad de implementar un proyecto de Big Data en una entidad, teniendo en cuenta factores como el volumen de información y los recursos disponibles.

RESUMEN

El propósito fundamental del presente escrito es realizar la recolección de información referente al concepto de Big Data, además se pretende establecer la viabilidad de integrar estos sistemas a nivel nacional analizando los resultados para los proyectos que se han desarrollado entorno a la problemática y comparando los medios y escenarios más eficientes para el uso del mismo.

Para cumplir con este propósito es necesario realizar la recolección de grandes niveles de información a nivel global y nacional, las implicaciones, beneficios y resultados de establecer sistemas integrando este tipo de manejo de información y de este modo concluir los medios óptimos para la implementación del sistema.

Con el análisis referente a la tecnología podremos estimar si la implementación de Big Data es necesaria y provechosa para todas las organizaciones o si por el contrario representa un manejo costoso de la información para una estructuración de una cantidad de datos que puede ser fácilmente referenciada por cualquier otro medio escatimando en el uso de los recursos, esto analizando el emprendimiento que ya distintos usuarios han realizado en cuanto a esta tecnología.

ABSTRACT

The main purpose of writing this is to perform the collection of information concerning the concept of Big Data also seeks to establish the feasibility of integrating these systems at the national level by analyzing the results for projects that have developed around the problem and comparing means and more efficient to use the same settings.

To fulfill this purpose it is necessary to collect high levels of information at the global and national level, the implications, benefits and results of establishing systems integrating this type of information management and thus conclude the optimal means for system implementation.

With the analysis relating to technology we can assess whether the implementation of Big Data is necessary and beneficial for all organizations or whether on the contrary represents a costly handling information for structuring an amount of data that can be easily referenced by any other means skimping on the use of resources, this enterprise already analyzing the different users in terms of this technology.

RESUMEN EJECUTIVO

La gran cantidad de información contenida en las redes sociales, bases de datos, centros de control de tránsito, los data center y demás focos de acumulación de información han alcanzado niveles bastante voluminosos, se estima que en 2013 cada día fueron creados cerca de 2,5 trillones de bytes de datos (2.5×10^{18})¹. Esto hace necesario el desarrollo de propuestas para el análisis de tanto volumen de información, Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales.

En Colombia es un tema que merece ser analizado debido al auge de las TICS en el país², el manejo de grandes cantidades de datos en diversas áreas como la salud la economía y la seguridad está ligado con su compleja captura, almacenamiento y visualización, este trabajo propone realizar un análisis de la viabilidad de la implementación de Big Data en el ámbito colombiano de las telecomunicaciones. Al finalizar se espera obtener un enfoque claro en el tratamiento de los grandes volúmenes de datos en Colombia, esto obedeciendo a la falta de información del tema tanto en la industria como en la academia, con una duración de ejecución de 4 meses en la que se realizará la investigación y el análisis de la viabilidad tanto económica como estructural de Big data en Colombia.

Objetivos Propuestos

Generales.

- Establecer la viabilidad de la implementación de redes Big Data en Colombia

Específicos.

- Realizar un estudio comparativo entre tecnologías disponibles en el mercado para manejo de grandes volúmenes de datos en Colombia.
- Generar una descripción detallada de los factores incidentes en la implementación de proyectos de Big Data con aras de determinar el costo de la iniciativa.
- Proponer un modelo de implementación de proyectos de Big Data orientado a PYME.

CAPITULO 1

ESTADO DEL ARTE

1.1 BIG DATA EN COLOMBIA

En una encuesta realizada por la EMC¹ en la cual participaron 183 tomadores de decisiones de las TI (Tecnologías de la información) los cuales analizaron las implicaciones y el potencial de implementar Big data en el país se hizo énfasis en la siguiente información:

Más del noventa por ciento de los empresarios creen que la implementación de esta tendencia en sus sistemas será útil en la toma de decisiones, la tercera parte de los encuestados consideran de por sí que esto también los ayudará a ser más competentes en el mercado, y dos terceras partes consideran que esta tecnología podría ayudarlos a prever ataques cibernéticos, no obstante el 42% de los encuestados sostienen que no está dentro de sus planes el implementar esta tecnología pues refieren que tanto la experiencia como el instinto de sus equipos ejecutivos son el factor fundamental para la toma de buenas decisiones para la compañía, esto deja en evidencia la falencia que hay en el país hacia la adopción de una cultura informática en contraste con otros países donde las decisiones se toman en base a estudios y mediciones.

A pesar de que muchos empresarios son conscientes de las implicaciones positivas que tiene el uso de Big data en el proceso de análisis y manejo de datos los mayores impedimentos para su popularización radican en el presupuesto, la falta de manejo del tema y la cultura como principal causal.

1.2 PROPUESTAS DE BIG DATA EN EL MINTIC

En Colombia la tendencia Big data es promovida por el MINTIC entidad que evidencia el desarrollo de esta herramienta para potenciar procesos industriales, estatales y educativos. En respuesta a la necesidad de Big data y de la importancia reconocida por el MINTIC en cuanto a los beneficios sociales, culturales, económicos y políticos que puede otorgar esta tendencia, el ministerio de las TIC ha creado centros de excelencia y apropiación (CEA) con el fin de centralizar el análisis de datos en sitios estratégicos para explotar al máximo las capacidades en el análisis de datos. Los objetivos de los CEA son focalizar el análisis de datos a sectores aprovechables y productivos que hacen referencia a innovaciones para las áreas de ciberseguridad, formulación de políticas públicas e internet de las cosas. El proyecto hace parte de la convocatoria 687 del MINTIC titulado CEA Big Data y Data Analytics el cual busca generar soluciones

respaldas por las TIC con aras de fortalecer el sector técnico del país, esta es una iniciativa que busca crear sectores especializados principalmente dirigidos a universidades certificadas, empresas líderes en Big data y empresas que no pertenezcan al sector TIC, este emprendimiento está vigente desde el 2014. Los denominados centros de excelencia buscan posicionar a Colombia como líder y referente en Big Data Analytics con beneficios de innovación nacional.

1.3 TENDENCIAS DE BIG DATA EN COLOMBIA PARA PYME

En Colombia la entrada de tecnologías e inversiones extranjeras es uno de los principales factores influyentes en el déficit de la economía nacional, es por esto que se hace fundamental analizar los referentes de Big Data actualmente operativos en Colombia:

1.3.1. BIG DATA EN ALMACENES, FIRMA SENSETA:

Senseta es una empresa que opera desde junio de 2013 en el país y se especializa en capturar una gran cantidad de datos por medio de sistemas innovadores haciendo uso de esa información con ayuda de la inteligencia artificial y la inteligencia de negocios con el fin de que los analistas puedan crear mejores estrategias y a su vez logren hallar tendencias que les den una ventaja competitiva en la industria, esta empresa busca también situar al país como uno de los referentes en el manejo de grandes volúmenes de datos en América latina².

Su operación consiste básicamente en la captura de información de sistemas, la herramienta de análisis de Senseta convierte estos registros en cifras reales, proporcionando estadísticas sobre conteo de personas, mapas de flujo, rutas de circulación, puntos críticos, zonas calientes y frías y tiempo de permanencia y demás factores lo cual conlleva a mejores decisiones en un negocio. El alcance de Senseta no se limita únicamente a un sector de la industria, por el contrario su tecnología es aplicada en agricultura, posicionamiento, telecomunicaciones, entre otros.

1.3.2. SENSETA TERRAINLOGICS:

Los modelos tradicionales para representar la realidad son simples y susceptibles al error humano, de tal modo que las decisiones que se realizan haciendo uso de esta información no son óptimas, por lo cual Senseta ha desarrollado una herramienta que provee la mayor precisión para describir la realidad con el mayor detalle posible haciendo uso de la recolección, análisis e interpretación de un gran volumen de datos.

1.3.3. TODOESTADISTICAS COLOMBIA:

En Colombia es necesario realizar una encuesta de satisfacción de los clientes para mejorar el servicio en las empresas, Todoestadísticas es una empresa que invita a sus usuarios a atreverse a dar el primer paso en campos donde otras empresas no ha incursionado para ser benefactores del uso del Big data en el análisis de datos y ver reflejados estos beneficios en el mercado. Todoestadísticas en Colombia es una firma estadística que cuenta con un grupo de expertos cuyo propósito es implementar y analizar el Big data en diversos tipos de empresas generando mayor rentabilidad como consecuencia del uso de Big data.

1.3.4. SAS EN COLOMBIA:

La modalidad de esta entidad consiste en predecir el comportamiento de los datos basados en citas y usos anteriores de los datos en cuestión, esto quiere decir que dependiendo la información y la toma de decisiones más concurrentes el aprendizaje automático o learning machine como se denomina al sistema es capaz de tomar decisiones congruentes y adecuadas a un sistema de entradas para optimizar su respuesta. La compañía de origen estadounidense y que opera en varios países del mundo implementa esta tecnología con el propósito de brindar una experiencia real y emocional del factor económico de las empresas para hacer uso de lo que es el recurso más importante en la actualidad, los datos. La implementación de esta tecnología permitiría la toma de decisiones en tiempo real de manera eficiente y precisa, por no citar aplicaciones como detección de fraudes financieros, establecimientos de nuevos modelos de precios o incluso resultados de búsquedas en la web.

1.3.5. DATA CENTER DE IBM EN COLOMBIA:

El proyecto consiste en un nuevo datacenter gestionado por IBM y que se ubica en Colombia, con esto se pretende hacer frente a las necesidades de una plataforma robusta y veloz que permita procesar, almacenar, transmitir y respaldar la información que se recoge desde diferentes fuentes como supermercados, bancos, entidades financieras y demás. El data center que se especializa en Big Data y Cloud se encuentra en un parque tecnológico situado a las afueras de Bogotá y su elaboración demandó una suma cercana a los 17 millones de dólares. Cuenta con una capacidad de almacenamiento de 3 petabytes y Velocidad de transmisión de 40 gigabytes.

CAPITULO 2

MARCO TEORICO

Big Data se plantea en el futuro de las tecnologías de la información y la comunicación no solamente porqué se adapta mejor a los cambios y evolución que está sufriendo la sociedad tanto en tecnología como en necesidades sino también por la gran aceptación que tiene entre las empresas. Esta aceptación es una realidad a dos niveles: a nivel de desarrollo, donde cada vez más compañías desarrollan nuevas tecnologías y apuestan más por las existentes como *Hadoop*³; y a nivel corporativo, donde las empresas empiezan a ver con más interés las nuevas posibilidades que aportan las soluciones Big Data, con el resultado del incremento de proyectos Big Data.

En Colombia el MINTIC se ha encargado de impulsar el conocimiento de este tópico para efectos de futuros proyectos tanto académicos como en la industria, con proyectos como el de BIG DATA EL FUTURO, y conferencias como el Encuentro Mundial de Big Data – Colombia 2015⁴ y la creación del Centro de Excelencia y Apropiación en Big Data y Data Analytics todos ellos encaminados a llevar una mejor acogida por el público escéptico en esta temática.

Este proyecto presenta una investigación orientada al ámbito colombiano que trata el Big Data como un tema poco explorado, tanto en áreas de seguridad como en el impacto económico que se puede lograr, de igual modo se propone un modelo de implementación orientado a PYMES en el cual se busca establecer la viabilidad de la implementación de la tecnología al mismo tiempo que se pretende alentar nuevas investigaciones que planteen la implementación de proyectos Big data.

2.1 ¿QUE ES BIG DATA?

Big data se describe como el conjunto de tendencias en avances tecnológicos encaminados a un nuevo enfoque en la manipulación de la información y toma de decisiones, este sistema puede emplearse para manipular datos estructurados, semiestructurados o no estructurados, entendiéndose por datos no estructurados aquellos no almacenados en una base de datos tradicional en tanto que los semiestructurados a pesar de carecer de una base de datos relacional (tiene la capacidad de interconectar o relacionar datos almacenados en distintas tablas) permiten una organización interna más sencilla de manipular⁵.

Esta tendencia busca manipular estas enormes cantidades de datos de una manera

eficiente solucionando las falencias de los actuales sistemas para procesar tales cantidades de datos. No obstante Big data no se refiere a una cantidad de datos específico aunque suele usarse usualmente cuando la cantidad de los datos merodean los petabytes y exabytes⁶, para darnos una idea de la cantidad de información que esto refiere es conveniente interpretar la cantidad de los datos en bytes.

$$\text{Gigabyte} = 10^9 = 1,000,000,000$$

$$\text{Terabyte} = 10^{12} = 1,000,000,000,000$$

$$\text{Petabyte} = 10^{15} = 1,000,000,000,000,000$$

$$\text{Exabyte} = 10^{18} = 1,000,000,000,000,000,000$$

Esto nos permite tener una idea de la gran cantidad de datos con la que opera el sistema, datos que según la Unión Europea se estiman en unos 1.700 nuevos billones de *bytes*⁷ por minuto, esta cantidad exorbitante de información hace necesario la implementación del Big data para el tratamiento de la información, a esto se debe sumar también que la información a manipular puede provenir de diversas fuentes como por ejemplo móviles, portátiles, sistemas de audio o video, sistemas de respuesta inmediata u otros, lo cual implica que su análisis sea aún más complejo debido a que además de las implicaciones del análisis de un gran volumen de información se le adiciona el factor velocidad al análisis de datos, nos permite entender que esta herramienta provee un medio para el análisis eficiente y veloz de una gran cantidad de información incluso para bases de datos no relacionales o parcialmente ordenadas como se enuncio anteriormente sin embargo es preciso destacar que los otros medios para el tratado de datos puede resultar eficiente para casos puntuales por lo que no sería coherente excluir otros medios por la existencia del Big data.

2.2 FUENTES DE DATOS:

Es inevitable el aumento considerable en la cantidad de datos que se manipulan diariamente en la humanidad teniendo en cuenta los constantes avances tecnológicos en los dispositivos que se conectan entre si e intercambian información constantemente, adicionado a esto el número de personas aumenta también con el paso del tiempo de una manera desmedida lo cual implica mayor flujo de datos dada la relación de usuarios. A este cúmulo de datos puede sumársele la información recolectada por entidades financieras como la información de sus clientes, proveedores, transacciones y demás, de manera similar ocurre en el sector público donde por ejemplo se mantienen índices de medición de la población para parámetros como su calidad de vida, economía, indicador laboral y en todo tipo de actividades que se nos puedan

presentar⁸, desde compartir una canción en la red hasta consultar información o incluso en aparatos comunes como neveras, impresoras entre otros elementos que emplean tecnologías de datos. Es evidente entonces que uno de los dispositivos que mayor flujo de información provee a nuestra base de datos serían los denominados Smartphone puesto que integran muchas de las tecnologías mencionadas anteriormente como cámaras, redes sociales, GPS, búsqueda de canciones y videos, entre otros y que cuentan con acceso directo a la red facilitando la producción de datos que haciendo una referencia puntual enmarcando la producción de datos a twitter y Facebook llegan a alcanzar los 12 Terabytes y 100 Petabytes respectivamente⁹.

En resumen puede decirse que la producción de datos de los Smartphone diariamente rodea los 2.5 quintillones de bytes diarios a nivel mundial, teniendo en cuenta que el equivalente de un quintillón es 10^{30} podemos evidenciar que esta es una cantidad de datos exorbitante. Sin embargo no es esta la cantidad de datos a la que debe hacer frente Big data únicamente sino que también debe tener en cuenta el crecimiento anual en el tráfico de datos que según una predicción realizada por CISCO¹⁰ establece un crecimiento del 78% al año y predice que el número de dispositivos conectados a la red superara eventualmente la cantidad de humanos en el planeta, lo cual implica que la cantidad de producción de datos anuales equivalga a los 130 Exabytes.

Vale la pena mencionar que no es únicamente el hombre quien provee la creación y manejo de datos pues existe también la comunicación máquina a máquina la cual crea volúmenes considerables de información en operaciones como la producción de información en el sensado de diversas variables como de temperatura, presión, electricidad, que se transfieren directamente de un dispositivo a otro.

2.3 TIPOS DE DATOS

Para hacer posible un buen manejo y análisis de la información a tratar es coherente establecer un arreglo o diferenciación entre los tipos de datos que se van a manipular, esto con el propósito de entender mejor su representación, aunque debe entenderse que esta clasificación puede variar por diversos factores incluyendo el avance tecnológico entre estos. Una clasificación del tipo de datos que trata Big data se referencian en la siguiente figura.

Web Y Redes Sociales	Máquina A Máquina	Transacciones De datos
Transmisiones de datos	Mediciones Inteligentes	Reclamos de Salud
Twitter's	Lecturas RFID	Telecomunicaciones y Detalles de llamadas

Publicaciones en Facebook	Señales GPS	Registros de facturas
Contenido Web		
Generados por humanos	Biométricos	
Grabación de llamadas en call - center	Reconocimiento Facial	
Email	Genética	
Registros médicos electrónicos		

Figura 1. Tipos de datos de Big Data

2.3.1. REDES SOCIALES:

Implica la información obtenida desde la web y desde redes sociales como Facebook, Twitter, LinkedIn, blogs, etc.

2.3.2. MÁQUINA A MÁQUINA (M2M):

M2M Describe la información generada entre dispositivos o máquinas, en ella se integra el uso de dispositivos como sensores o medidores que capturan algún evento (velocidad, temperatura, presión, cambios climatológicos, etc) los cuales pueden ser transferidos a través de redes alámbricas, inalámbricas o híbridas hasta otras aplicaciones que transforman los datos recolectados en información significativa para el usuario.

2.3.3.- TRANSACCIONES:

Incluye facturación, en telecomunicaciones registros detallados de las llamadas. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

2.3.4. BIOMÉTRICA:

Datos como 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.

2.3.5. GENERADAS POR LOS HUMANOS:

Los humanos contribuimos en gran medida con la producción de datos generados desde diversas fuentes como por ejemplo la información que guarda un call center al establecer una llamada telefónica, correos, documentos electrónicos, notas de voz, etc.

2.4 LA IMPLEMENTACIÓN REVOLUCIONARIA DE BIG DATA

Hemos visto con antelación que la inclinación en el uso de las tecnologías Big Data refieren el almacenamiento y procesamiento de grandes volúmenes de datos, lo cual se evidencia en el contexto colombiano, no obstante se puede apreciar que el uso de estas tecnologías se aplican en diversas instancias del desarrollo nacional, se aprecian entonces empresas que incursionan en el uso del Big Data, lo cual les permite tomar mejores decisiones en su entorno de una manera veloz y manipulando tipos de datos variados integrando así diversos sistemas de negocios y optimizando los procesos.

Encontramos en el panorama nacional entidades que realizan una implementación del sistema de Big Data de una forma directa, analizando únicamente los beneficios del medio, sin embargo es necesario tener presente que no en todos los sistemas es este el método más eficiente para el manejo de información, esto refiriéndose a la cantidad de datos y el coste que pueda representar para la entidad teniendo en cuenta que existen otros procesos para el manejo de datos que se acoplan a volúmenes menores de información obteniendo resultados similares y más económicos.

De manera similar existe la implementación del Big Data desde un método evolutivo, en este el equipamiento del método lo realizan empresas con trayectoria en el manejo de datos y que cuentan con una inteligencia empresarial sostenible y con un manejo de la información más elaborado, lo cual implica el conocimiento de herramientas de análisis de datos, warehouse y demás elementos previos a la introducción de la tecnología, un claro ejemplo es el data center de IBM estructurado en Colombia dado que esta empresa tiene trayectoria en el manejo de datos y han evolucionado su método hasta estructurar plataformas Big Data eficientes, esto se ha logrado gracias al procesamiento de datos desde las plataformas anteriores que se han ido introduciendo de manera gradual en estos nuevos sistemas lo cual permite el análisis de datos de una manera más veloz, variada y amplia y además reducen el umbral de costos en la modernización de la tecnología.

A pesar de que la mayoría de las implementaciones de la tecnología Big Data se enfocan en el procesamiento de información almacenada previamente o de manera

continúa en bases de datos también existen aplicaciones innovadoras en el mercado, tales como las que provee la firma Senseta, empresa que ha logrado convertir registros de sensores, GPS y diversas maniobras en cifras manipulables por los sistemas de manera innovadora y reduciendo el error humano en los procesos de cálculo.

VENTAJAS:

- Rendimiento. Los modelos ahora se nutren con muchos más datos y más diversificados.
- Volumen. Se multiplica el volumen de información que se obtiene, gracias a Big Data.
- Ahorro. Tanto en tiempo de implantación, como en costes, ya que se mantiene el BI existente.

INCONVENIENTES:

- La velocidad de extremo a extremo siempre estará limitada por el entorno actual de BI (que es mucho más lento).
- El nivel de percepción no es tan amplio, ya que fallaría la granularidad por el BI existente.
- Esta solución no es definitiva, porque llegará un momento en que el BI no tenga capacidad para hacer frente a los requerimientos informativos de la organización y se tenga que optar por hacer el traslado completo de los datos al exterior de la empresa.

2.5. LA ALTERNATIVA HÍBRIDA A LA IMPLANTACIÓN DE BIG DATA

En la alternativa híbrida del uso de las tecnologías se alterna el sistema dependiendo de la finalidad que se persigue y de los requerimientos del trabajo a realizar, entonces pueden utilizarse los sistemas BI previos para el análisis de datos e información no muy complejas y para el trato de información que requiera mayor capacidad hacerse uso del Big Data, para hacerlo posible es necesario que los datos del data warehouse se almacenen en un motor de análisis, en otras palabras un datamart, de este modo se consigue un análisis avanzado, mayor precisión en el manejo de datos, actualizaciones en tiempo real o análisis predictivos, haciendo uso de los sistemas tradicionales.

VENTAJAS:

- Precisión: los sistemas ahora se nutren con una información estructurada, cuando antes sólo podían acceder a datos no estructurados.
- Big Analytics. O lo que es lo mismo, posibilidad de hacer un análisis predictivo, creando un modelo estadístico con todos los datos e identificando las relaciones causales y las correlaciones; apoyándonos también en herramientas avanzadas de visualización.
- El internet de las cosas: que permite conocer en tiempo real todo lo que está sucediendo en cualquier parte y en relación a cualquier asunto.
- Rentabilidad. En definitiva, esta opción permite ofrecer mejores servicios que, claro está, también pueden cobrarse.

INCONVENIENTES:

- Resistencia: de las empresas menos evolucionadas, tecnológicamente hablando, a la implantación de un modelo de este tipo.
- Dificultad: de quienes están acostumbrados a limitarse a un tipo de análisis simplemente descriptivo, a la hora de lanzarse a las nuevas posibilidades.

2.6 PANORAMA MUNDIAL

2.6.1. ÁMBITO INTERNACIONAL

Como una tendencia que no para, así podría definirse Big Data en el mundo pues cada país aún por desarrollado que parezca tiene innumerables campos de desarrollo para esta tecnología creada por IBM, puede verse también en el gran número de personas conectadas a internet o utilizando alguno de los servicios que ofrece, a gran escala podría verse por algunos de los indicadores del banco mundial, como el de los usuarios de internet por cada 100 personas entre el 2011-2015, donde es Islandia el país con más usuarios de internet por cada 100 personas con un total de 98,2 seguido de Dinamarca con 96,2, por su parte Colombia ocupa un puesto muy relegado con 52,6 usuarios de internet por cada 100 habitantes, sin embargo cabe denotar su crecimiento

de 12 puntos porcentuales en comparación al año 2011, Este indicador es imprescindible para analizar el ámbito internacional pues como elemento primordial para la extracción de datos está el internet y sus servicios derivados este indicador dará a priori un elemento de decisión a la hora de implementar redes de análisis de datos.

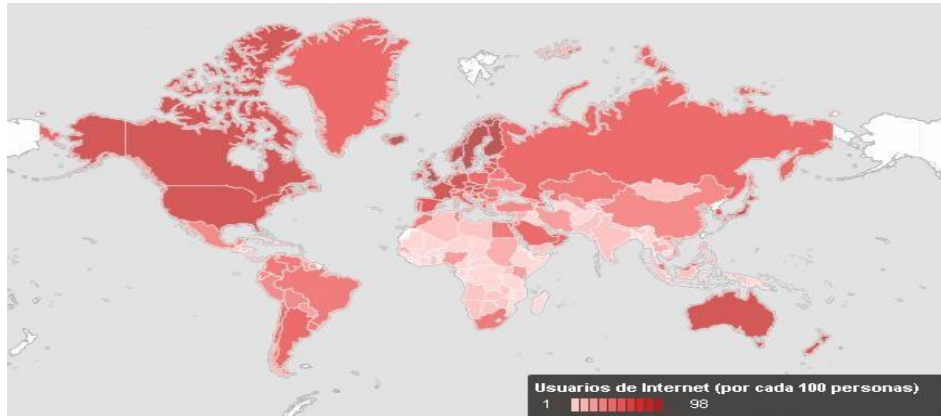


Figura N° 2. Usuarios de internet por cada 100 personas (Banco Mundial)¹¹

Existen en este momento en el mercado plataformas potentes para el análisis de datos que están siendo usadas y moviendo una cantidad de recursos tanto financieros como de personal que antes no era si quiera posible pensar, casos como Data integrador de Oracle, y las plataformas de SAS que son los más usados a nivel mundial presentan una variedad de datos al explorar y experimentar para interacciones personalizadas con el usuario más precisas, y una anticipación más rápida y ajustada a las conductas y las situaciones cambiantes naturales en la conducta humana. Es decir la toma de decisiones en efectos prácticos. Luego podría surgir la siguiente pregunta ¿Por qué es importante?, la importancia del Big Data reside en que los datos se pueden extraer de cualquier fuente y permiten el análisis de información para encontrar respuestas que facilitan:

- Reducir costos
- Reducir tiempos
- Desarrollar nuevos productos y optimizar las ofertas
- Tomar decisiones inteligentes

Cuando se combina Big Data con High powered analytics¹², Se pueden llevar a cabo tareas relacionadas con el negocio, tales como:

- Determinar la raíz de fallas o defectos, en tiempo real.

- Generar cupones en el punto de venta basados en los hábitos de consumo de tu consumidor.
- Recalcular riesgos de portafolios enteros en minutos.
- Detectar comportamiento fraudulento antes de que afecte a tu organización.

Hacia mediados 2009, el mundo experimentó una pandemia de gripe A, llamada gripe porcina o H1N1. El sitio web Google Flu Trends¹³ fue capaz de predecirla gracias a los resultados de las búsquedas. Flu Trends usa los datos de las búsquedas de los usuarios que contienen Influenza (Síntomas parecidos a la enfermedad de la gripe) y los agrega según ubicación y fecha, y es capaz de predecir la actividad de la gripe hasta con dos semanas de antelación más que los sistemas tradicionales. Más concretamente en Nueva Zelanda cruzaron los datos de Google Flu Trends con datos existentes de los sistemas de salud nacionales, y comprobaron que estaban alineados. Los gráficos mostraron una correlación con las búsquedas de los síntomas de la influenza y la extensión de la pandemia en el país. Los países con sistemas de predicción poco desarrollados pueden beneficiarse de una predicción fiable y pública para abastecer a su población de las medidas de seguridad oportunas.

El contexto internacional entonces ofrece una visión holística de lo que vendrá en cuanto a tendencias durante los siguientes años, en especial para zonas como Latinoamérica y el caribe donde la inmersión de estas es poca.

2.6.2. ÁMBITO EUROPEO

El Barcelona Supercomputing Center (BSC) y Microsoft Research Center han iniciado el Proyecto Aloja, una iniciativa conjunta que busca conseguir una optimización automática en despliegues de Hadoop, en infraestructuras distintas. El primer objetivo de este proyecto es de desarrollar un estudio sistemático del rendimiento de ejecuciones en Hadoop, en diferentes configuraciones de hardware y software, desplegadas tanto en servidores locales como en infraestructuras cloud. Los resultados que se obtengan estarán disponibles a través de un portal del Barcelona Supercomputing Center. La finalidad que se busca es que las empresas o personas que trabajan en el sector de la investigación, despliegue o promocionen soluciones Hadoop y puedan beneficiarse de la base de datos, a la que se irán incorporando resultados de rendimiento y guías de configuración.

El proyecto Aloja se encuentra en una fase de planificación preliminar y de preparación cuenta con un presupuesto para el primer año de alrededor de 1,1 millones de dólares. El equipo está dirigido por Rob Reinauer (Microsoft) y David Carrera (BSC).



Figura N° 3 Laboratorio Arte Alameda, Proyecto Aloja Barcelona España ¹⁴

Desde el año de su creación, en 2008, el BSC-Microsoft Research Center ha llevado a cabo diversos proyectos de investigación sobre las interacciones de software en procesadores multi-núcleo y en arquitecturas paralelas de procesamiento masivo, y en este momento está incubando nuevos proyectos relacionados con el cloud computing y el Big Data.

La iniciativa de Europa para el 2020¹⁵ creada por la agenda digital Europa en 2013 busca transformar las industrias de servicios de Europa mediante la generación de una amplia gama de productos y servicios de información innovadores; aumentar la productividad de todos los sectores de la economía a través de la mejora de la inteligencia empresarial; abordar de manera más adecuada a muchos de los retos a los que se enfrentan la sociedad europea, mejorar la investigación y acelerar la innovación; lograr reducciones de costos a través de servicios más personalizados y aumentar la eficiencia en el sector público.

2.6.3. ÁMBITO ASIÁTICO

El departamento de estrategias de IT del Gobierno japonés está elaborando una reforma de la legislación relativa a las tecnologías de la información, como parte de la “Nueva estrategia de IT”¹⁶ de la Administración del primer ministro Abe. Esta nueva estrategia, que se basa en un documento llamado “Manifiesto para la creación de la nación más avanzada del mundo en IT”¹⁷, abarca medidas concernientes al uso privado de los

datos de los organismos públicos y al del Big data, sistemas de gestión de gigantescas bases de datos con información obtenida de fuentes como los teléfonos móviles. La gestión de este tipo de datos presenta dos grandes problemas: la seguridad y la protección de la privacidad. Para abordarlos, el departamento de Estrategias de IT está preparando una serie de propuestas legales, como la enmienda de la ley de protección de la información personal.

El sector japonés que es uno de los grandes administradores mundiales de tecnología se ha enfocado en la protección y el manejo ético de los datos recaudados y consignados en sus gigantescas bases de datos pues esta tendencia lleva alrededor de 7 años gestándose en el país nipón. El departamento de estrategias de IT constituyó el comité de análisis sobre los datos Personales, en el que los expertos en la materia debatieron la reforma, y en junio de 2015 presentaron un “Paquete de revisiones del sistema regulador del uso de datos personales”. En ese primer borrador se clasificó en dos categorías la información que se encuentra en la llamada “zona gris” entre la información personal como el nombre y la dirección y la no personal: “datos de baja especificidad personal” (nombre provisional), información que puede utilizarse sin el permiso del individuo si se mantiene el anonimato, y “datos secundarios de identificación personal”, información como la ID de los terminales móviles y los datos de reconocimiento facial. Esta definición de conceptos se pretendía plasmar en la ley para determinar cómo debían tratarse los distintos tipos de información por parte de los organismos oficiales y las empresas pero luego de diversos debates tanto de oposición como apoyo a la reforma se llegaron a establecer los siguientes tópicos principales:

- La creación de un organismo externo para supervisar la legislación
- La reforma de la legislación requiere el consenso de la ciudadanía
- Definir la posición del organismo supervisor externo
- Buscar un equilibrio con las regulaciones internacionales

2.6.4. ÁMBITO AMERICANO

América en este caso en concreto Estados Unidos tiene que ser vista como uno de los principales afluentes de información del mundo, sus tendencias no solo marcan la moda y el avance en algunas capas de la tecnología sino también la diversificación de servicios prestados pues es un país pluricultural, diversos centros de investigación desde la universidad de Harvard hasta el Instituto de astronomía estadounidense y la NASA han incursionado en las tendencias del Big data, por obvias razones son entidades con remarcable hermetismo en sus adelantos, sin embargo los grandes desarrollos que están a simple vista son en lo que se centrará el ámbito americano de esta investigación, pues su parte de políticas y desarrollo de legislación de igual manera

es un tema con poca información al momento.

Casos como el de Macy's que es uno de los comercios minoristas más importantes de los Estados Unidos, que destaca por su e-commerce¹⁸. Utilizando la tecnología de SAS Institute ha conseguido mejorar sus ingresos y la experiencia del usuario. Gracias a la velocidad de análisis y los informes obtenidos con esta nueva tecnología, han reducido en dólares el gasto anual de analítica. Macy's sabe hoy perfectamente el impacto de sus 'newsletters' y notificaciones y conoce mejor a los clientes más satisfechos, lo que les gusta y lo que no.

Hoy, el uso de estos datos, les permite segmentar al máximo sus envíos, de manera que envían menos emails, pero con mucho más impacto y han conseguido reducir las de suscripciones hasta en un 20%. Gracias a la utilización de un algoritmo y al control de la demanda y el inventario, pueden lanzar ofertas cruzadas, ajustar precios y hacer rebajas casi en tiempo real para sus 73 millones de artículos a la venta. La Información financiera aquí presentada fue obtenida mediante el análisis de graficas en el informe de E-tech de MIT¹⁹.

Tras su primer mandato, el presidente de los EEUU, Barack Obama, decidió utilizar Big Data para su reelección en 2012. Un centenar de personas trabajaron en el departamento de analítica de la campaña. 50 estaban fijos en las oficinas centrales, otros 30 se movilizaron a lo largo y ancho de las distintas sedes del país, y 20 estaban única y exclusivamente centrados en la interpretación de los datos recibidos. Tras un primer análisis, los esfuerzos de la campaña se enfocaron en tres aspectos:

- Registro (recoger datos de los votantes convencidos)
- Persuasión (dirigirse a los dudosos de una forma eficaz)
- Voto del electorado (asegurarse de que los partidarios fueran a ejercer el voto sí o sí)

Y, por primera vez, los tres equipos más importantes de las campañas electorales: el de campo, el digital y el de comunicación, trabajaron con una estrategia unificada con los respectivos datos de cada uno. El motor de todo, la plataforma inteligente utilizada fue HP Vertica ²⁰. Entre las acciones más efectivas que permitía esta plataforma estaban: recoger datos a pie de campo y realizar un feedback muy rápido vía notificaciones email por parte del equipo online (se mejoraba en tiempo y eficiencia); o detectar los nichos en los que funcionaría mejor la publicidad en TV cruzando datos de los votantes con otros demográficos, audiencias, precios de publicidad, programas. Con su analítica, el equipo de Obama optimizó la comunicación y mejoró la respuesta del electorado afín, permitiendo no malgastar recursos, tiempo y dinero en los votantes que no eran partidarios de su partido.

2.7 ANÁLISIS ÁMBITO NACIONAL

El escenario colombiano en el desarrollo de Big Data es un fenómeno emergente que lleva tras de sí un compendio de tecnología, infraestructura, inversión social e inversión del sector privado pues el financiamiento de este tipo de mega proyectos es la base fundamental para su éxito posterior debido a que no es un retroactivo inmediato es decir para ver la eficiencia y el margen de productividad se necesita que el sistema lleve en funcionamiento un tiempo prudente, tanto entre recolección, análisis e interpretación de datos que es la estructura general para poder hablar de Big Data, luego de tomar como referencia los más representativos avances en el mundo en esta temática se puede trasponer algunos al ámbito nacional y dilucidar un poco el rumbo de las futuras investigaciones en el país.

Es el caso del banco Bancolombia que utilizando el análisis de datos ofrecido por SAS latino América se convirtió en un caso de éxito, con 140 años de historia, Bancolombia no solo ha logrado consolidarse como el grupo financiero más importante de Colombia y uno de los más grandes de Latinoamérica, sino que también se posiciona como una de las compañías colombianas que más aprovechan los datos en la actualidad para tomar decisiones y diferenciarse en el mercado. ¿Cuándo ampliar el cupo de una tarjeta de crédito, ofrecer un crédito o llamar la atención de un cliente por haberse pasado en el pago de alguna obligación? Todas estas son decisiones que los bancos están tomando cada vez con mayor inteligencia y que les están permitiendo ofrecer mejores productos al mercado sin arriesgarse.

La analítica permitió generar informes como el de saldos diarios, el cual se ha convertido en una de las principales fuentes de información para la toma actual de decisiones en todo el Banco, ya que es utilizado por más de 600 personas que conforman las direcciones de Riesgo de Crédito, Vicepresidencia de Empresas y Gobierno, Vicepresidencia de Personas y Pymes y gerencias de conciliación con clientes, desde las primeras horas del día.

En el informe del primer semestre de 2015 en la entidad bancaria revela que tuvo un crecimiento porcentual del 2.9% y el valor de la unidad invertida presento el siguiente comportamiento

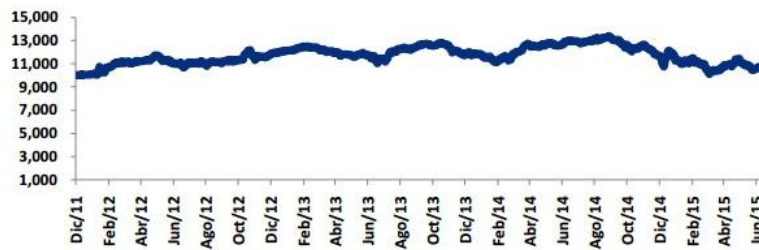


Figura N° 4 Evolución del valor de la unidad Bancolombia²¹

Lo que evidencia que el banco ha tenido el crecimiento económico notable desde la utilización de data 'Analytics y Big'²² Data, Latinoamérica es precisamente una de las regiones del mundo que más campo de acción tiene para usar esta herramienta tecnológica. De acuerdo con un estudio publicado por la revista Forbes y realizado por Dell a empresas del ramo de las Tecnologías de Información, 40 por ciento de las empresas en Norteamérica buscan activamente adoptar o expandir el uso de Big data, 50 por ciento en Asia, 35 por ciento en Europa y 30 por ciento en Latinoamérica.

Data Analytics es una rama de Big Data utilizada en el solo proceso de análisis de datos, a diferencia de otras tecnologías esta solamente se centra en resultados estadísticos

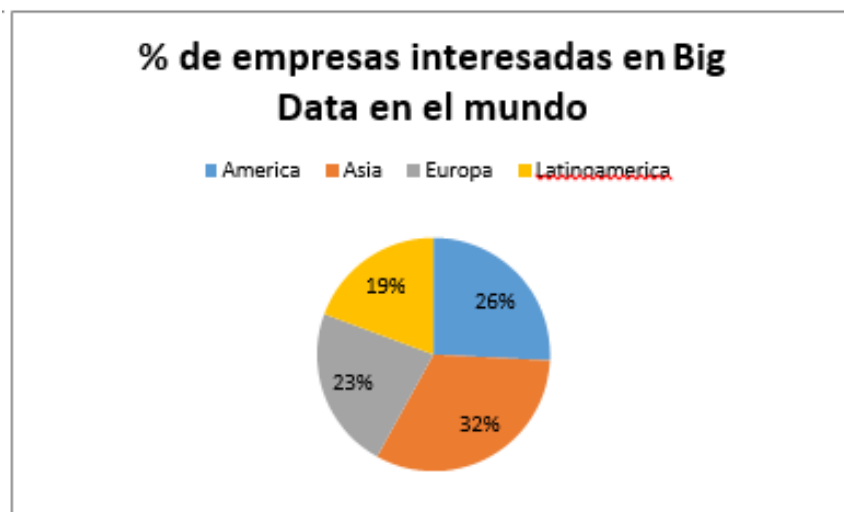


Figura N° 5 Porcentaje de empresas interesadas en Big Data en el mundo

Colombia tiene la oportunidad de convertirse en un referente regional en la materia, por

lo que desde el Gobierno, la academia y la empresa privada están empezando a nacer alianzas que permitan al país capitalizar el análisis de datos en sectores estratégicos. Una de las tácticas es la creación de centros especializados en análisis de datos. Para ello el Ministerio de las TIC y Colciencias abrieron una convocatoria dirigida a universidades certificadas en alta calidad, empresas líderes en Big Data y empresas ancla que no pertenezcan al sector TIC.

La convocatoria estuvo vigente hasta el 28 de septiembre de 2015 y las entidades que se beneficien de la misma se tendrán que encargar de construir un centro de investigación aplicada que cuente con grupos de investigación, programas de formación y que promueva el emprendimiento. Además, este centro de investigación debe desarrollar proyectos demostrativos del uso de Big Data como herramienta para empresas y para el sector público.

CAPITULO 3

COMPONENTES DE UNA PLATAFORMA BIG DATA

Dado que el Big data implementa procesos que son utilizados cuando los métodos habituales para el tratamiento de datos supera la capacidad del software se hace necesario implementar componentes más efectivos, de este modo podemos hablar del uso de Hadoop para el manejo de datos.

3.1. HADOOP

Es una infraestructura digital con licencia apache creada bajo código abierto que permite la programación utilizando java, el propósito de esta tecnología es generar un motor de búsqueda potente y eficaz haciendo uso de tecnologías inspiradas en google. La importancia de esta estructura radica en que permite el paralelismo en sus procesos segmentando la información y dividiéndola en n equipos de cómputo permitiendo así el análisis de Petabytes de datos. De este modo siendo Hadoop uno de los hilos de Big data es conveniente enunciar de manera detallada su funcionamiento, dicho lo anterior Hadoop se segmenta en tres partes principales:

- *Hadoop Distributed File System (HDFS)*
- *Hadoop MapReduce*
- *Hadoop Common.*

En resumen es un sistema de código abierto que se utiliza para almacenar, procesar y analizar grandes volúmenes de datos; cientos de terabytes, petabytes o incluso más.

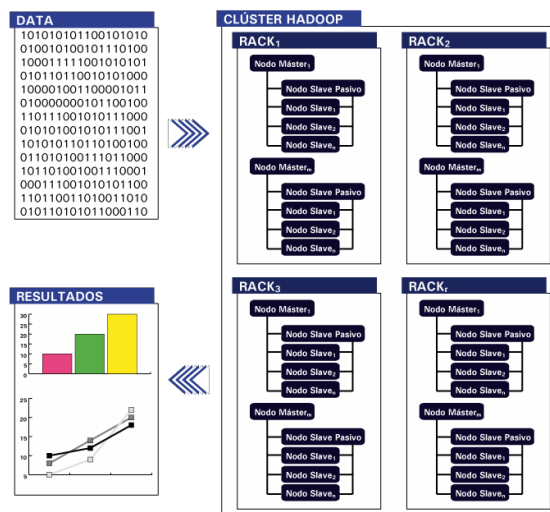


Figura N° 6 Funcionamiento de Hadoop.

3.2. FUNCIONAMIENTO HADOOP

Hadoop funciona como un sistema de archivos cuyo propósito es recolectar la mayor cantidad de información posible, el funcionamiento consiste en el establecimiento de clúster HDFS. Cada nodo establece bloques de datos sobre la red usando protocolos de bloqueo para el HDFS. El sistema de archivos se comunica a través del protocolo TCP/IP mientras que para la comunicación entre clientes se utiliza el RPC, el cual permite a un ordenador ejecutar código en otro sin preocuparse por los protocolos de comunicación. La estructura del HDFS busca almacenar estas grandes cantidades de información en múltiples máquinas lo cual permite que no requieran almacenamiento RAID (sistemas de almacenamientos de datos en discos duros, SSD) entre sí. La información se distribuye en tres nodos, dos de los cuales se sitúan en un mismo rack y el restante en uno distinto., en este sistema los nodos de datos son capaces de intercambiar datos entre ellos, mover copias y conservar la replicación de datos para salvaguardar la información.

HDFS no cumple totalmente con POSIX porque los 43 requerimientos de un sistema de archivos POSIX difieren de los objetivos de una aplicación Hadoop, porque el objetivo no es tanto cumplir los estándares POSIX sino la máxima eficacia y rendimiento de datos. HDFS fue diseñado para gestionar archivos muy grandes. HDFS no proporciona alta disponibilidad.

3.3. MAPREDUCE

MapReduce podría definirse como un framework, es decir, representa una arquitectura de Software (lenguaje de programación), muy usado en la programación de funciones de alto nivel, como se puede ver en la figura 7 algunas de las funciones con más importancia son:

`map(list[], oper)` aplica la operación 'oper' a la lista 'list[]', retornando una nueva lista cuyos elementos han sido operados, individualmente, por 'oper'. Por ejemplo, si disponemos de la lista `list[1,2,3]` y de la operación suma $X = X + 1$ y realizamos la operación `map(list, suma)` el resultado de la misma será `[2,3,4]`.

`fold (list[], oper)` aplica la operación 'oper' a la lista `list`, retornando un elemento producto de la operación de los elementos de `list` entre sí. Por ejemplo, si disponemos de la lista `list[1,2,3]` y de la operación suma $(x: xs) = x + \text{suma}(xs)$ y realizamos la operación `fold(lista, suma)`, el resultado de la misma será 6.

Debido a la posibilidad de que la operación no sea asociativa, los lenguajes de programación ofrecen, normalmente, dos operaciones: una “hacia la derecha”, o foldr y otra “hacia la izquierda”, o foldl. En general, en el área de la computación distribuida Map se utiliza para fraccionar una operación compleja entre varios nodos y Fold/Reduce para recoger los resultados y unificarlos. Por su parte, los frameworks MapReduce toman la base de las operaciones mencionadas anteriormente para crear una operación genérica y más compleja, cuyo funcionamiento es realmente útil para las bases de datos NoSQL en vez de usarse sobre listas de valores unidimensionales, ésta toma como parámetros entrantes una lista de tuplas²³ de tipo (clave, valor) y devuelve una lista de valores. Entre las operaciones map (distribuida) y reduce (normalmente localizada) se genera una lista de tuplas (clave, valor) con valores temporales, de las que reduce y filtra solamente las que tengan una determinada clave. En definitiva, MapReduce es fundamental en las bases de datos NoSQL para permitir la utilización de funciones de agregación de datos, ya que al carecer de esquema son mucho más complicadas que en las bases de datos relacionales clásicas.

Un ejemplo de uso de MapReduce es MongoDB que es un sistema de base de datos NoSQL orientado a documentos, desarrollado bajo el concepto de código abierto y nacida en 2007. En vez de guardar los datos en tablas como se hace en las bases de datos relacionales, MongoDB guarda estructuras de datos en documentos tipo JSON²⁴ con un esquema dinámico (MongoDB llama ese formato BSON), haciendo que la integración de los datos en ciertas aplicaciones sea más fácil y rápida. Esta base de datos es altamente utilizada en las industrias, por ejemplo son muy utilizadas por MTV Network, Foursquare.

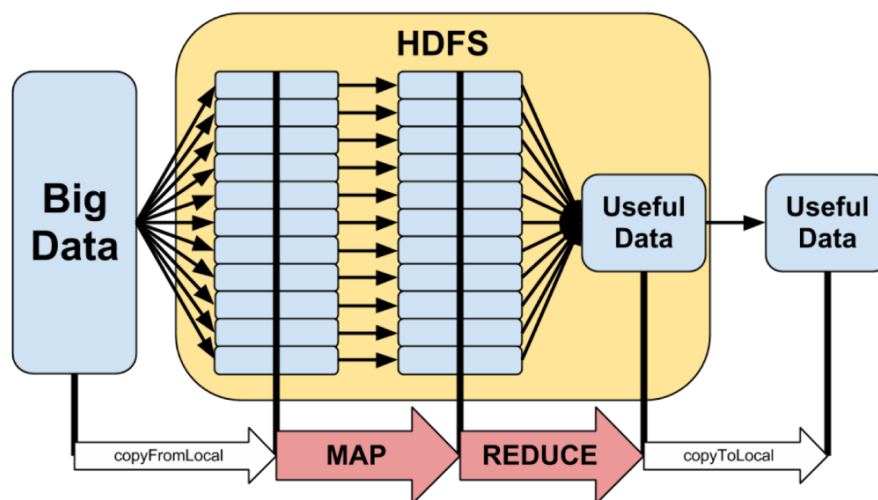


Figura N° 7 Estructura de MapReduce.²⁵

CAPITULO 4

BIG DATA COLOMBIA

Los sistemas financieros, de salud, agrícolas y demás en el mundo están enfrentando importantes transformaciones que impactan no solo sus modelos de negocio sino la forma de relacionarse con los consumidores.

4.1 ALIANZA CAOBA (Centro de alianza y aplicación en Big Data y Data Analytics):

Uno de los más importantes problemas en el país es la existencia de barreras que impiden la rápida apropiación de las tecnologías necesarias para hacer uso masivo y efectivo de las diversas herramientas de Big Data. La especialización, variedad y velocidad con la que avanzan las tecnologías asociadas al dominio del problema, sumado a las inversiones requeridas por la industria para implementar estrategias de Big Data, y a la falta de suficiente recurso humano altamente capacitado, hacen que la implementación de políticas de Big Data por parte de las medianas empresas, no vaya a la velocidad que requiere el país y no tenga el cubrimiento necesario para impulsar la industria nacional.

El anuncio sobre la iniciativa CAOBA³⁷ fue hecho en el lanzamiento del Centro de Excelencia en Big Data y Data Analytics, la primera alianza publico privada que tiene múltiples participantes, entre ellos la DNP, como líder política en Big Data para el estado Colombiano. CAOBA es única en su tipo, siendo el primer proyecto de este tipo en el país y el único de Latinoamérica que involucra academias, el sector privado y el sector público y el cual busca obtener soluciones a problemas reales en diferentes sectores de la industria.

La alianza CAOBA, que operará como un Centro de Excelencia y Apropiación y cuya entidad ejecutora es la Pontificia Universidad Javeriana, contará con la participación del Grupo Nutresa, Bancolombia y DNP, Universidad Icesi, Universidad Eafit, Universidad de Los Andes, IBM, EMC, SAS y Clúster Creativ y tiene como objetivo desarrollar un sistema de capacidades analíticas en el país para contribuir a la competitividad y al posicionamiento de la cultura en innovación e investigación aplicada basada en el uso de grandes volúmenes de datos; y donde el papel del Departamento Nacional de Planeación (DNP) constituye la creación de una política reguladora del Big Data y Data Analytics en el país y pretende afectar positivamente varias áreas de la economía nacional:

Agua: Análisis para optimizar el recaudo y generar esquemas de ahorro.

Movilidad: Análisis de los patrones de desplazamiento de la población, para mejorar la movilidad y la seguridad ciudadana.

Educación: Evaluación de la rentabilidad de la inversión en educación por análisis en cadena.

Residuos sólidos: Análisis del volumen y tipo de desechos en una zona y el aprovechamiento que se les está dando.

4.2 BBVA Y BIG DATA:

El objetivo principal es capturar valor de los datos que tiene el banco. Dado que el banco posee datos de diversas índoles como sus clientes, obligaciones y demás, lo que se busca en primera instancia es centralizar esos datos, para poder analizarlos y extraer el conocimiento necesario o la información relevante, lo cual es posible debido al manejo en la variedad de datos que permite el Big Data, esta información se traduce en nuevos y mejores servicios de las entidades financieras para con sus clientes.

Por ello, para el banco BBVA la aplicación más importante de Big Data es extraer información de los clientes a partir de lo que hacen, esto implica que la información que se extrae para realizar las encuestas y estadísticas en aras de mejorar el servicio provienen de una fuente más fidedigna que una encuesta, pues implica seguir el patrón de uso de los servicios por parte de cada cliente para determinar que movimientos ejecuta, que tipo de vida tiene y de este modo generar un catálogo de actividades y servicios concordes a su actividad. En este sentido, la parte de inteligencia comercial para poder dar ese mejor servicio es fundamental. Y, además, ese conocimiento de los clientes nos permite evaluar mejor al cliente de cara a productos que tenemos como puede ser dar créditos o prevenir el fraude por citar algunos ejemplos.

4.3 DAVIVIENDA:

Según IDC³⁸, el mercado de soluciones de Big Data y analítica en América Latina representó 661 millones de dólares en 2014. Se espera que la adopción de las soluciones de analítica y Big Data se incremente como resultado de la madurez del ecosistema de Hadoop y el incremento en el uso de herramientas y aplicaciones analíticas como las ofrecidas por empresas especializadas en Big Data analytics, como SAS. En Colombia, el banco Davivienda ha tomado técnicas emergentes de Big Data

para analizar los comportamientos típicos de grupos de clientes similares y controlar o rechazar usos atípicos en productos como el de DaviPlata, un ejemplo claro acerca del uso efectivo de la seguridad en sistemas de gran volumen.

4.4 BIG DATA Y EL SECTOR DE LA SALUD: EL FUTURO DE LA SANIDAD

Los pacientes, las clínicas, los hospitales tienen cantidades masivas de datos clínicos, en formatos escritos en papel o electrónicos pero que permanecen sin utilizar por la dificultad e imposibilidad material de “digerirlos” de forma efectiva, por muy buenos deseos que pueda tener el equipo sanitario. Lo cierto es que esta dificultad puede tener consecuencias tanto en el control de los gastos médicos como en la mejora de las tasas de mortalidad.

Este es el futuro de la salud, según ha publicado la prestigiosa revista Forbes: los big data representan una oportunidad para los innovadores y todos los que se preocupan por la salud, aumentan substancialmente la posibilidad de obtener información más efectiva de los datos y menores tasas de mortalidad de los pacientes.

4.5 EL ESTADO DE LA CUESTIÓN

El modelo sanitario y en general el sector de la salud, es uno de los sectores donde Big Data está teniendo mayor impacto en la actualidad y donde sus aplicaciones crecerán de un modo espectacular, tanto para el área médica, como también para las áreas de análisis de datos (historias médicas, análisis clínicos...), la gestión de centros de salud, la administración hospitalaria, la documentación científica (generación, almacenamiento y explotación).

De acuerdo al estudio de 2011 del Kinsey Global³⁸ Instituto sobre Big Data, sin lugar a dudas uno de los más referenciados en la Web, calcula que las aplicaciones de Big Data en el campo de la salud podrían suponer un beneficio de 250.000 millones de euros al sector público europeo y unos 300.000 millones de dólares al sector de sanidad de los EEUU.

Otro informe también de gran impacto es, “Big Data in digital Health” de la Fundación Rock Health³⁹ que analiza la situación actual y el potencial del Big Data en el mundo de la salud. El informe utiliza los datos y la información obtenidos en entrevistas con emprendedores e inversores y calcula que se pueden producir importantes ahorros en el sector sanitario. Ahorros que calcula entre 325 y 525 millones de dólares que divide de la siguiente forma: 1. 25-50 millones de dólares en la mejora en la coordinación de la

atención al ciudadano; 2. la lucha contra los fraudes y los abusos; 3. la lucha contra las ineficiencias administrativas y clínicas. Según las conclusiones del informe hay seis vías mediante las cuales Big Data puede cambiar la atención sanitaria, apoyo a la investigación):

1. Transformación de datos en información.
2. Apoyo al autocuidado de las personas.
3. Apoyo a los proveedores de cuidados médicos.
4. Aumento del conocimiento y concienciación del estado de salud.
5. Agrupamiento de los datos para expandir el ecosistema.

4.6 AREA DE LA SALUD - CONVERSATORIO SOBRE BIG DATA Y GESTIÓN DE LA SALUD EN LAS CIUDADES

El Observatorio de Sociedad, Gobierno y Tecnologías de la Información y Comunicaciones de la Universidad Externado de Colombia realizará un conversatorio en el que presentará las tendencias jurídicas, nacionales e internacionales, vinculadas con el uso de Big Data en las ciudades, específicamente en el sector salud.

El objetivo de este evento es promover un diálogo sobre el alcance técnico y jurídico y la relevancia del Big Data para la gestión de la salud en las ciudades. Este evento tuvo lugar el 14 de marzo de 2016 en Colombia,

4.7 SALUD - ALIANZA CAOBA

La alianza CAOBA también incursiona en el área de la salud además de los ámbitos anteriormente mencionados, con esto se pretende optimizar el tiempo de respuesta ante procesos de rutina y se tiene la expectativa ante un uso más eficiente y veloz en el registro de los millones de datos existentes para optimizar el uso de los recursos del Estado. En cifras estas mejoras se evidencian en los 2,7 billones anuales de pesos que la EPS invierte en administración, cifra que puede ser optimizada implementando soluciones análisis con los más de 300 millones de registros de datos anuales que maneja el sistema de salud

CAPITULO 5

VIABILIDAD DE BIG DATA EN COLOMBIA

Para hablar de viabilidad de un tema tan amplio como es Big Data, deben primero sentarse los parámetros sobre los cuales se piensa analizar el proyecto, es decir en qué área específica de producción, servicios o entretenimiento se piensa hacer el diagnostico de viabilidad, en este capítulo se piensa tomar en consideración los grandes sectores a tener en cuenta en Colombia y se enfocará en el sector de las telecomunicaciones pues es el principal campo de interés para esta monografía , los sectores a tener en cuenta son :

- Sector de producción:
Agrícola
Industrial
De comercio y/o publicitario
- Sector de comunicaciones:
Internet
Televisión
Telefonía
- Sector Bancario
- Sector Salud

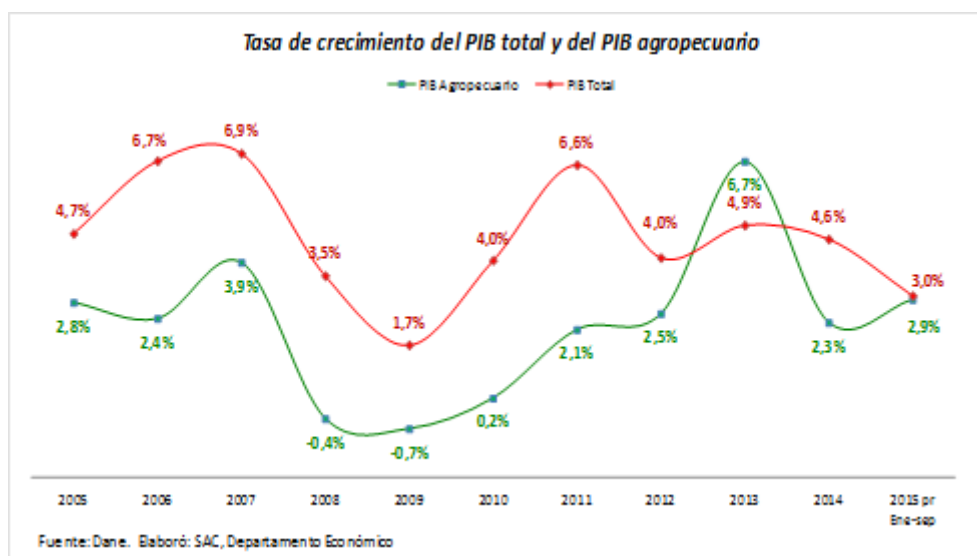
Se pretende medir la viabilidad con indicadores como porcentajes de crecimiento del sector TIC, penetración y la incidencia que tiene en sector TIC en áreas de interés en Colombia, el cubrimiento de internet, uso de TICS en PYMES en Colombia, el posicionamiento a nivel internacional que tienen los proyectos realizados en el país, las iniciativas presentadas por el DANE y el MINTIC entre otros, con el fin de estimar un nivel de aceptación y porcentajes de crecimiento para las PYMES con favorabilidad de implantar Big Data.

5.1 SECTOR AGRÍCOLA

El sector agrícola en Colombia ha sufrido notables cambios en las últimas décadas, primero la inclusión de nuevos productos extranjeros han afectado en gran medida el

precio de venta de los alimentos producidos por los agricultores colombianos, sin embargo durante los primeros nueve meses del año 2015, el PIB agropecuario creció 2.9%, comparado con el mismo periodo de 2014, lo que indica un incremento de 0.1 puntos porcentuales por encima del nivel reportado un año atrás (2.8%) e inferior al total acumulado de la economía nacional que reporta un 3.0%. El crecimiento se atribuye al buen comportamiento de la producción cafetera que de enero a septiembre de 2015 acumula un alza de 11.5% y de renglones pecuarios como porcicultura (11.8%) y avicultura de carne (6.0%). (Los datos que aquí se citan son obtenidos de reportes oficiales del MINTIC y el DANE adjuntos en la sección de bibliografía) ²⁶

Sin embargo, descontando el café se puede observar que el sector agrícola tuvo solo un 1.1% de crecimiento. Con excepción de arroz (12.9%), palma de aceite (9.9%), cacao (15.8%) y frutales (7.1), que mostraron importantes incrementos en sus niveles de producción; los demás cultivos tuvieron comportamientos moderados e incluso descensos significativos, de acuerdo con la información gremial. Este es el caso de cultivos como maíz (-18.4%), sorgo (-36.7%), cebada (-41,3%), trigo (-42,9%), algodón (-22,8%), tabaco (-13,1%), soya (-8.8%), frijol (-8.2%) y hortalizas (-4.4%).



La actividad agropecuaria colombiana en 2016 podría estar creciendo en su conjunto entre un 1% y 2%, como resultado de la dinámica de la actividad cafetera y, en menor proporción, del ingreso a producción de áreas de cultivos de ciclo largo. Infortunadamente, los factores que determinan la inversión en el sector se tornan poco favorables en el corto plazo, debido al entorno macroeconómico local. Fedesarrollo calcula un crecimiento del PIB de 2,7% para Colombia en 2016. En el panorama mundial se observa un reducido crecimiento de la economía estadounidense; la zona

Europea comienza a salir lentamente de su estancamiento y Japón se recupera levemente.

Por otra parte con un presupuesto total de 4.300 millones de pesos ²⁷, el MinTIC y su aliado Colciencias, abre la convocatoria 'TIC y Agro' en primer trimestre de 2016 para cofinanciar proyectos orientados al desarrollo de soluciones tecnológicas innovadoras que beneficien a las diferentes comunidades organizadas de productores agropecuarios, es decir: cooperativas, fundaciones, gremios, asociaciones y Organizaciones No Gubernamentales (ONG) que trabajen de manera directa en actividades agrícolas relacionadas con el desarrollo rural.

El valor máximo a financiar por cada proyecto será de 400 millones de pesos, con una contrapartida del 5% al 15% dependiendo del tamaño de cada empresa de TI: 5% de cofinanciación para las micro, 10% para las pequeñas, y 15% en el caso de las medianas. Las soluciones que las compañías de TIC desarrollen serán de su propiedad, como una estrategia del gobierno para fortalecer a la Industria TI, y también como una fórmula para incentivar que haya más aplicaciones y herramientas digitales para el agro colombiano.

Las participantes de la convocatoria 'TIC y Agro' podrán hacer sus propuestas teniendo en cuenta líneas temáticas relacionadas a las cadenas de producción agrícola: sistemas de información técnica, de mercado, de propiedad de la tierra, datos geo - referenciados, información de clima, desarrollos para llevar información a los campesinos a través de terminales inteligentes y de gama baja, bancarización, educación, sistemas de información de precios, distribución y productos, entre otros , de esta manera se denota un amplio espectro de incursión que podrían tener los proyectos de Big Data en este sector aunque por su nivel de inversión de capital podría verse estancado en sus primeras fases pues hay que recordar que son iniciativas de un coste considerable en las primeras etapas con un nivel de retorno de inversión bastante bajo.

5.2 SECTOR INDUSTRIAL

El sector industrial en Colombia por su parte en un informe realizado por el MINCIT en abril de 2016 la industria del país creció un 8.4% ²⁸, con lo que se puede afirmar que el buen momento por el que pasa la industria no es fortuito. Es fruto de los esfuerzos que vienen haciendo los empresarios por incorporar en sus negocios la productividad y la innovación para crear valor agregado, diversificarse y acceder a nuevos mercados el Programa de Transformación Productiva (PTP), desde el 2011 y hasta abril de este año,

se ha destinado aproximadamente \$54 mil millones, para el desarrollo de iniciativas en pro del fortalecimiento de los 20 sectores que lo conforman, entre los que se encuentran manufacturas, agroindustriales y de servicios. También las acciones de iNNpulsa Colombia han sido determinantes para los resultados que hoy se muestran. Desde que fue creada en 2002, ha asignado recursos de cofinanciación por \$206 mil millones para apoyar y fomentar el emprendimiento, la innovación y el fortalecimiento empresarial

De acuerdo con el reporte del Dane en abril de 2015²⁹, 29 de las 39 actividades industriales estudiadas en la encuesta presentaron una variación positiva. Sin tener en cuenta la refinación de petróleo, la producción industrial reportó un aumento del 5,2%. Las que más aportaron a esa dinámica fueron bebidas (14,6%), hierro y acero (18,4%), jabones y detergentes, perfumes y preparados de tocador (10,4%). De acuerdo con las cifras preliminares del DANE en la Encuesta Anual de Servicios, el mayor crecimiento del consumo intermedio en el sector TIC fue entre 2012 y 2013, con un incremento del 11,3%. El crecimiento presentado desde el 2009 del consumo intermedio del sector TIC, comprueba que cada año el sector TIC se comporta como un sector transversal en la economía colombiana, por lo que tiene influencia en el crecimiento de los demás sectores.

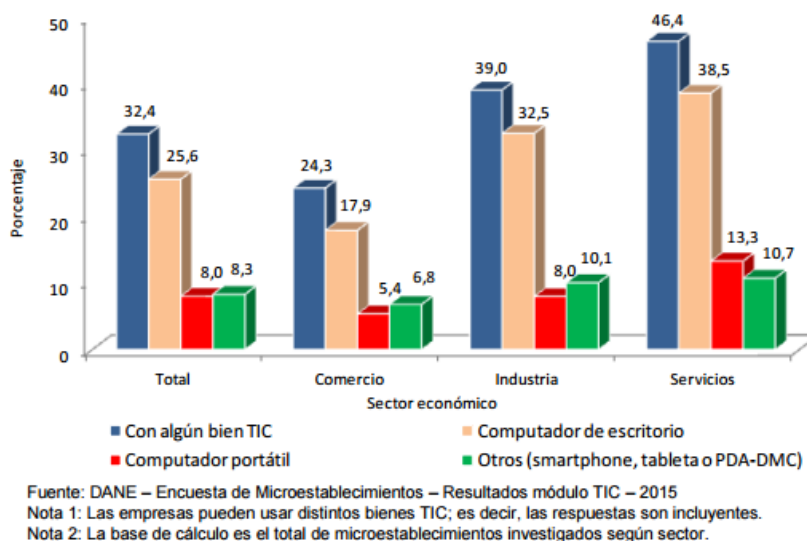
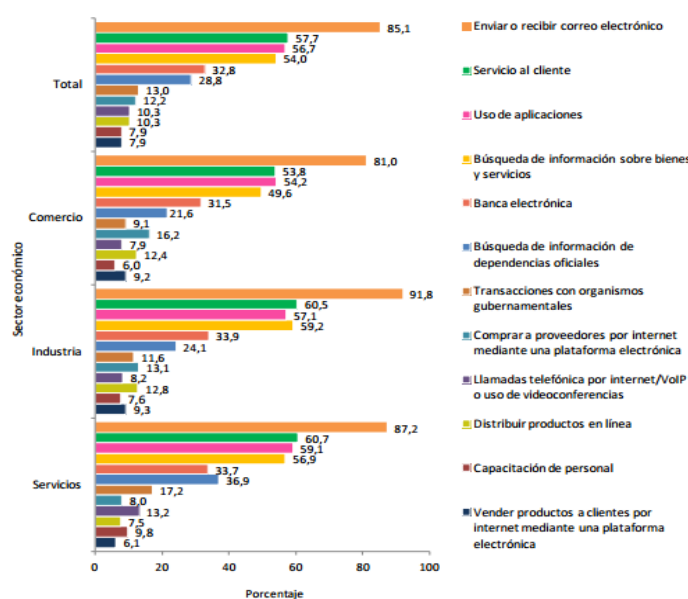


Fig.8 Proporción de micro establecimientos que usan algún servicio TIC

2015.* Fuente: Indicadores Básicos de Tenencia y Uso de Tecnologías de la Información y Comunicación – TIC en Microestablecimientos
 2015-DANE-1 de abril de 2016³⁰

Los resultados de evolución (PANEL) 2014 - 2015 se refieren a un total de 33.785 establecimientos de industria, comercio y servicios³¹ ubicados en las 24 ciudades principales y sus áreas metropolitanas. En 2015, el 32,4% de los microestablecimientos investigados usaba algún bien TIC; 39,0% en el sector industrial, 24,3% en comercio y 46,4% en servicios. En 2015, el 26,2% de los microestablecimientos investigados tuvo acceso o utilizó el servicio de Internet; por sectores, 32,2% en industria, 18,8% en comercio y 39,0% en servicios. En 2015, el 22,9% del personal ocupado en los microestablecimientos investigados tuvo acceso o utilizó Internet; 22,3% en industria, 17,5% en comercio y 31,9% en servicios. En 2015, el 5,9% de los microestablecimientos investigados tenía página web o presencia en un sitio web; 8,3% en industria, 4,6% en comercio y 7,7% en servicios.

De los microestablecimientos que no usaron bienes TIC, el 90,6% argumentó que no los necesitaron y por sector, 90,5% en industria, 90,6% en comercio y 90,8% en servicios. El costo y no saber usarlos, fueron las siguientes razones que adujeron los microestablecimientos que no usaron estos bienes TIC, con 6,0% y 3,3%, respectivamente.



Fuente: DANE – Encuesta de Microestablecimientos – Resultados módulo TIC – 2015

Nota 1: Las empresas pueden usar Internet para distintas actividades; es decir, la respuesta es incluyente.

Nota 2: La base de cálculo es el total de microestablecimientos que usaron Internet según sector.

Fig 9. Proporción de microestablecimientos que tuvo acceso o usó Internet, según actividad de uso

* Fuente: Indicadores Básicos de Tenencia y Uso de Tecnologías de la Información y Comunicación – TIC en Microestablecimientos 2015-DANE-1 de abril de 2016

Del total de micro establecimientos investigados que tuvo acceso o usó Internet en 2015, el 85,1% lo hizo para enviar o recibir correo electrónico; 57,7% lo empleó para servicio al cliente, 56,7% lo empleó para uso de aplicaciones, y 54,0% para búsqueda de información sobre bienes y servicios como se puede ver en la figura. De los micro establecimientos industriales que usaron Internet, el 91,8% lo hizo para enviar o recibir correo electrónico; 60,5% lo empleó para servicio al cliente, 59,2% para búsqueda de información sobre bienes y servicios y 57,1% lo empleó para uso de aplicaciones.

En los microestablecimientos comerciales, el uso de Internet para enviar o recibir correo electrónico registró el mayor porcentaje con 81,0% de los micro establecimientos que usaron Internet, seguido por el uso de aplicaciones y el servicio al cliente, con 54,2% y 53,8%, respectivamente. Finalmente, para el sector de servicios, el uso de Internet para enviar o recibir correo electrónico fue la actividad que registró el mayor porcentaje con 87,2% de los microestablecimientos que usaron Internet, seguido por servicio al cliente y uso de aplicaciones con 60,7% y 59,1%, respectivamente.

Estos indicadores son un aliciente bastante bueno para el análisis de la viabilidad de Big Data en Colombia pues se ve con claridad que los micro establecimientos o PYMES , tienen una favorabilidad aceptable a la hora de usar las TIC en sus entidades pues se ve con buenos ojos el cambio hacia nuevas tecnologías que optimicen sus procesos y les generen una mayor rentabilidad en el mercado , pues como se presentó en apartados anteriores el sector TIC es un sector que puede considerarse transversal a todas las actividades económicas del país pero genera un cambio positivo en todas las que tiene incidencia sin embargo las principales razones por las que no son aceptadas las TIC están entre el costo y el desconocimiento y podría radicar allí el frente sobre el cual se puede trabajar en aras de conseguir una mayor penetración de mercado.

5.3 SECTOR DE COMERCIO Y/O PUBLICITARIO

Es evidente que uno de los principales campos de acción de Big Data está en la publicidad, pues tras la recolección de datos y su interpretación se pasa a intervenir en el usuario final y este es el objetivo principal del sector publicitario pues conocer las necesidades y gustos particulares de los clientes es un avance formidable a la hora de vender o promocionar un producto, en el ámbito colombiano como se evidencia en la figura 10 , La actividad Correo y Telecomunicaciones se encuentra en el noveno lugar entre las 53 actividades que le aportan al crecimiento del PIB. La gráfica 10 muestra el peso relativo de las 15 actividades económicas que hacen parte del PIB 2014 – 3T, y comprenden el 28% del total de actividades (53), las cuales están generando el 68% del PIB



Fig.10 Ranking de actividades económicas en Colombia que aportan al crecimiento económico anual del país.

* Fuente: Panorama TIC informe oficial
Marzo de 2015-Mintic

Esto indica que el tránsito financiero es favorable para las empresas en el sector de publicidad pues su sector se ubica en el de comunicaciones, la publicidad está cambiando de una forma asombrosa no solo en Colombia sino en el mundo. Las viejas reglas del negocio han dado un giro de 180 grados. El objetivo es comprar audiencias, no espacios. Ahí entra la precisión del Big Data, la potencia de la máquina y el buen hacer de los desarrolladores.

La publicidad programática es el proceso por el que los anunciantes compran publicidad mediante procesos de pujas en tiempo real (RTB o Real-time Bidding) a través de distintos actores que facilitan el negocio: los Ad Exchanges, casas de subastas que ofrecen impresiones al mejor postor; Demand Side Platforms (DSP's), proveedores de

tecnología que permiten la optimización del precio gracias al Big Data (manejan grandes cantidades de datos sobre los usuarios: perfiles, historial de conversión, target , entre otros); los Data Partners, los que proporcionan los datos; los Trading Desk, el equipo del anunciante que trabaja directamente con las casas de subastas; y los Sell Side Platforms (SSP's), los que maximizan el rendimiento de los espacios publicitarios para los medios. Un estudio de IDC señala que el crecimiento esperado para la publicidad en tiempo real es del 59% anual hasta 2016 y que su volumen de negocio llegará al 40% de toda la publicidad móvil y en display en 2017.

Por su parte El Departamento Administrativo Nacional de Estadística DANE diseña y desarrolla el Sistema Estadístico Nacional SEN, con el fin de garantizar el cumplimiento de las buenas prácticas estadísticas y la generación de estadísticas de calidad para la toma de decisiones en el país. En tal sentido, el DANE implementó el Archivo Nacional de Datos ANDA un catálogo en el que los usuarios pueden explorar, buscar, comparar, solicitar acceso y descargar información relacionada con censos, encuestas por muestreo y uso estadístico de registros administrativos. El ANDA contiene metadatos de operaciones estadísticas que producen el DANE y otras entidades del SEN. Asimismo, algunos de los archivos tienen disponibles microdatos de acceso público.

La información publicada en el ANDA está documentada bajo los estándares internacionales DDI ³² y Dublin Core³³; esto fortalece la transparencia, comparabilidad, calidad, confianza y credibilidad de las estadísticas que produce el SEN, entidad bajo la cual se deberán regir las políticas del uso de datos como el caso de la publicidad programática.

5.3.1 INTERNET

El internet ha cobrado amplia importancia para la humanidad en las últimas décadas, afectando a la mayor parte de la población en diversos ámbitos. No obstante, los efectos de Internet sobre variables económicas y sociales dependen significativamente del tipo de uso que le sea dado a esta herramienta. De la misma manera, la utilización de internet depende de la disponibilidad de contenidos y aplicaciones de utilidad para los usuarios finales.

No obstante, según ITU ³⁴ en la actualidad existe una brecha de usuarios de internet influenciada por género, educación, ingreso y edad. Adicionalmente, existen grandes diferencias entre áreas rurales y urbanas, las cuales pueden estar dadas no solo por la diferencia en niveles educativos e ingreso, sino también por niveles deficientes de infraestructura presentes con mayor énfasis en las áreas rurales. Para países de América Latina, la evolución del número de usuarios de internet por cada 100 habitantes

muestra una tendencia creciente y sostenida en los países de la región.

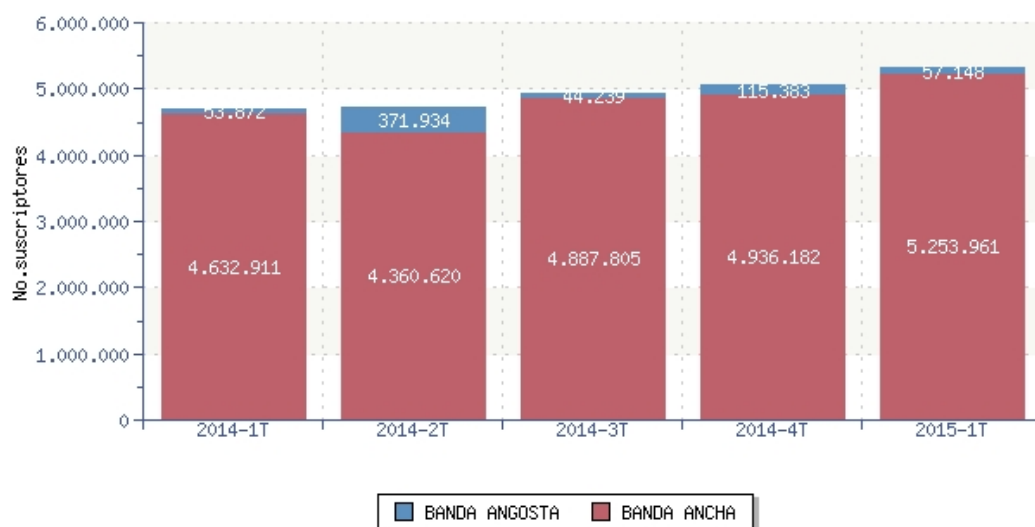


Fig.11 Suscriptores de internet por ancho de banda. Fuente: Mintic informe avance TIC 2015-1t³⁵

5.4 SECTOR FINANCIERO

Después de un periodo inicial donde el Big Data era considerado algo opcional para la gran mayoría de los usuarios, actualmente se considera ampliamente por su valor y eficiencia, empezando a formar parte del día a día de las entidades. El reto en el desarrollo de la tecnología en cualquier sector deriva la gestión de datos aplicada los tres conceptos principales de Big data: volumen, velocidad y variedad en el manejo de datos, conceptos de los que se ha visto beneficiado de manera evidente el sector financiero. Sin embargo existen muchos inconvenientes que impiden desarrollar Big Data de una manera confiable, muchos de ellos implican la falta de conocimiento del tema y la ausencia del personal calificado para la ejecución de las labores, otros que implican factores económicos o incluso tiempo, entre otros, los cuales se ilustran en la siguiente relación:

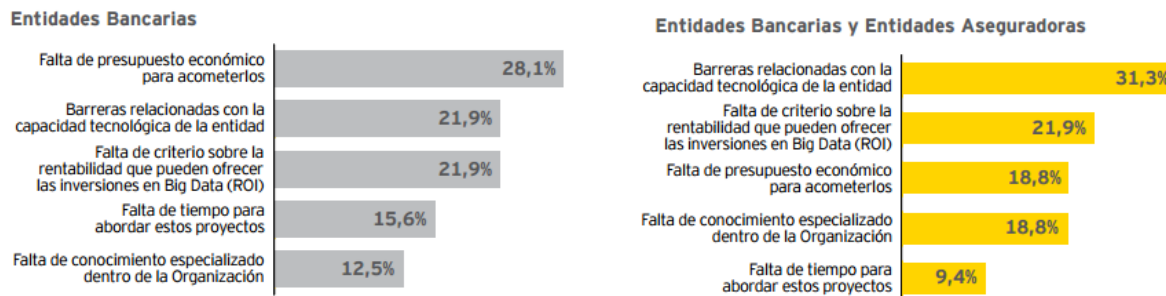


Figura.12 Análisis EY, FrontQuery y Teradata, Barreras en el sector financiero

Sin embargo Teradata resume las barreras para la implementación de big data en el sector financiero en seis, estas barreras se establecieron como conclusiones derivadas de investigaciones en las que participan entidades como EY, FronteryQuery y Teradata.

- **Generar conocimiento a partir de los datos:** Cuando se realiza la recolección de los datos, muchos de estos resultan ser información no deseada o no relevante, razón por la cual es conveniente transformar los datos en el tipo de información que la organización necesita para tomar las mejores decisiones de manera más veloz al eliminar el análisis de datos innecesarios, una vez seleccionada la información influyente es necesario verificar la veracidad de la misma, las herramientas analíticas a emplear y generar un análisis de aptitudes y talentos necesarios en el área.
- **Decidir qué es lo que se quiere lograr con los datos:** Para que todo sistema funcione debe tenerse claro desde un principio la meta y el curso a seguir, los problemas de la organización, las hipótesis de solución y las herramientas a usar para desarrollar la idea, en Teradata recomiendan hacer uso de los datos ya existentes en la empresa para comenzar con el análisis de la información, generando confianza y credibilidad en la organización.
- **Definir modos de medir los resultados:** El éxito en la implementación de Big Data se hace palpable realizando las mediciones de rentabilidad que genera, la reducción de costos y el incremento de ingresos, es por eso que se debe superar la barrera del desconocimiento verificando resultados cuantificables, lo cual permitirá enfocar la iniciativa en los campos más vulnerables en la organización.
- **Gestionar los riesgos:** Minimizar el riesgo ante aspectos de privacidad de los datos y ciberseguridad es fundamental para que cualquier sistema que maneje una cantidad exorbitante de información, haciendo énfasis en la adaptabilidad y responsabilidad, este ítem constituye el riesgo más importante de todos.

- **Prepararse para la transformación:** Dentro de cualquier organización los cambios resultan ser ordenes de sus directrices, por lo cual resulta conveniente que desde allí mismo se fomente el cambio y la selección de los talentos internos en la empresa, promoviendo la adopción de mejores sistemas, tecnologías y hábitos que permitan una integración de todas las funciones estratégicas de la organización (Dirección Financiera, Dirección de Sistemas, Dirección de Marketing, etc.), lo cual supone una transformación importante y que afecta a toda la organización.
- **Ir un paso más allá:** Para potenciar los resultados de la tecnología deben definirse planes de acción que se puedan llevar a cabo en plazos determinados, evaluando los resultados obtenidos e identificando su impacto positivo en la entidad.

Adicionalmente a estas seis barreras que deben superarse para obtener un resultado positivo con la tecnología, las firmas EY, FrontQuery y Teradata ejecutaron una encuesta en donde recopilan cuales serían los caracteres más influyentes que representan oportunidades estratégicas, basada en inteligencia de clientes enfocada al sector financiero:

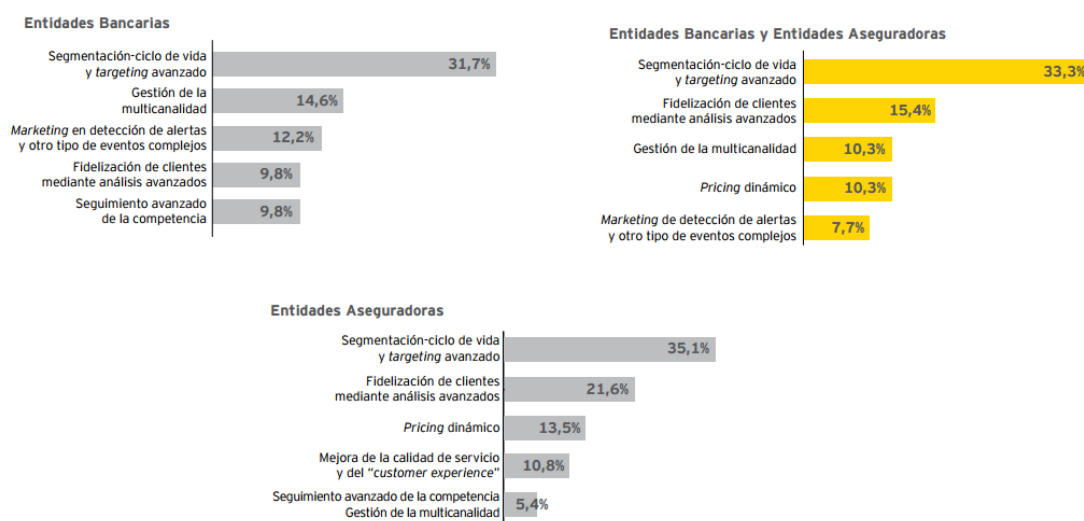


Figura . Análisis EY, FrontQuery y Teradata, oportunidades estrategicas en el sector financiero

La segmentación-ciclo de vida y targeting avanzado son las principales oportunidades en el sector.

La integración de información provenientes de fuentes internas y externas arraigadas a la Variedad en la manipulación de datos como dimensión relevante de Big Data es, tanto

para los bancos como para las entidades aseguradoras que han respondido la encuesta, la segunda gran ventaja a obtener.

5.4.1 ASPECTOS PROMETEDORES EN EL SECTOR FINANCIERO:

La tendencia Big Data permite a las organizaciones aprovechar nuevas fuentes de datos para mejorar la calidad en la toma de decisiones, posibilitando un análisis de datos y generación de información útil en tiempo real más veloz, lo cual se traduce en resultados de negocio, tanto por el incremento de eficiencias como por aumento en los ingresos.

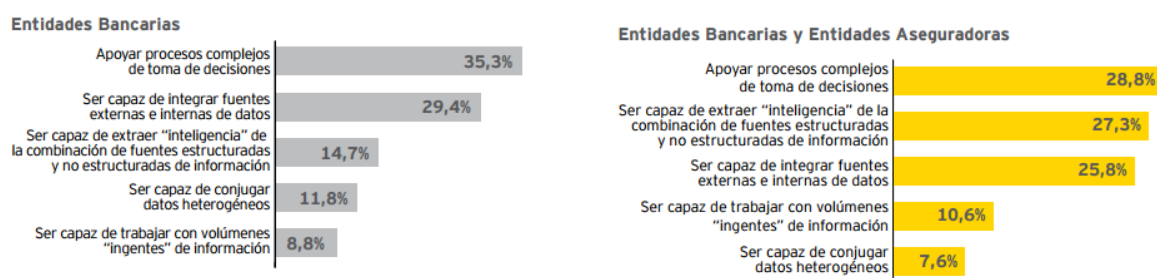


Figura.13 Análisis EY, FrontQuery y Teradata, aspectos prometedores en el sector financiero

Según una encuesta realizada por EY, FrontQuery y Teradata, los aspectos más prometedores en el sector financiero son el apoyo en la toma de decisiones inteligentes basándose en la experiencia del cliente en tiempo real y la integración de datos provenientes de fuentes externas e internas como se aprecia en el gráfico, lo cual encamina los esfuerzos de las compañías a explotar estas áreas fuertes para sacar el mayor provecho en Big Data³⁶.

Este análisis previo en cuanto a oportunidades, barreras y aspectos por explotar permiten posteriormente verificar la viabilidad de las tecnologías Big Data en el sector financiero, y pueden prevenir el desgaste innecesario en incursiones improductivas para la tecnología en áreas que no la requieren, es de esta manera como se aprecia la implementación gradual de proyectos financieros que requieren del uso de Big data, permitiendo así la participación de entidades bancarias en proyectos de Big Data en alianza con otras áreas de la economía.

CAPITULO 6

ESTUDIO COMPARATIVO DE TECNOLOGIAS OFRECIDAS EN EL MERCADO PARA BIG DATA EN COLOMBIA

Para el siguiente estudio comparativo se tuvieron en cuenta las tecnologías con mayor impacto no solo en Colombia sino en Latinoamérica, debido a que algunas tienen un alcance mundial forman parte del estudio, se tuvieron en cuenta parámetros de velocidad, variedad y volumen según la definición de meta datos para Big, en adición se tuvo en cuenta el sector económico sobre el cual se tiene mayor influencia pues es un factor clave en el momento de tomar una decisión sobre la tecnología a implementar, decidir pues la experiencia y la trazabilidad de los proyectos dependen en un gran margen de los casos de éxito anteriores.

	SAS
VELOCIDAD	Sas permite a los analistas para crear aplicaciones que pueden procesar y analizar rápidamente los volúmenes de eventos que fluyen de forma continua mediante la aplicación de streaming de análisis / consultas continuas. Los programadores pueden crear aplicaciones con la API C ++ de modelado o la capa de modelado XML que se incluye con el producto. Con la última versión, SAS ESP 3.1, los modelos ESP se pueden desarrollar utilizando la nueva SAS ESP Studio, un entorno de diseño de arrastrar y soltar gráfica para construir y probar sus modelos de ESP y tomar ventaja de todas caso la potencia de procesamiento de SAS ESP. Streaming de fuentes de entrada, de las operaciones, sensores, twitter, etc , Con una velocidad en tiempo real.

VARIEDAD	<p>La oferta avanzadas Gestión de datos SAS incluye todos los productos que se indican a continuación. Asteriscos (*) indican los productos de software que se incluyen sólo con la oferta de gestión de datos SAS avanzada.</p> <p>Base SAS and SAS Studio DataFlux Authentication Server* DataFlux Data Management Server DataFlux Data Management Studio DataFlux Secure DataFlux Web Studio and Server* SAS BI Mobile * SAS Business Data Network* SAS Business Rules Manager (optional)* SAS Data Integration Studio SAS Data Quality Server SAS Data Remediation SAS Integration Technologies SAS Job Monitor SAS LASR Analytic Server or SAS LASR Analytic Distributed Mode Server* SAS Lineage* SAS Management Console SAS/CONNECT</p>
VOLUMEN	SAS ofrece una capacidad de almacenamiento de 2,5 Hasta los 3 Petabytes
ARQUITECTURA	<p>La modalidad de esta entidad consiste en predecir el comportamiento de los datos basados en citas y usos anteriores de los datos en cuestión, esto quiere decir que dependiendo la información y la toma de decisiones mas concurrentes el aprendizaje automático o learning machine como se denomina al sistema es capaz de tomar decisiones congruentes y adecuadas a un sistema de entradas para optimizar su respuesta.</p>
COSTO	800-1000 USD
SECTOR ECONÓMICO	<p>La principal industria en la cual se basan los casos de éxito de SAS, están el sector de la banca, seguido de comunicaciones , servicios estatales y por ultimo seguros, además tiene incursiones en hotelería y turismo , y oil & gas</p>

VIGENCIA EN EL MERCADO	<p>SAS Lanza su primer software de procesamiento de datos en 1997, principalmente para el análisis de datos agrícolas, SAS alguna vez significó "Statistical Analysis System" ("Sistema de Análisis Estadístico") y se inició en la Universidad Estatal de Carolina del Norte como un proyecto para analizar investigación agrícola. Conforme aumentó la demanda de este software, SAS fue fundada en 1976 para ayudar a toda clase de clientes (desde compañías farmacéuticas y bancos hasta entidades académicas y de gobierno). En adición ha crecido premios como una de las mejores 8 compañías para trabajar en Europa además de galardones por los aportes hechos a la ciencia y la estadística. Actualmente es una de las multinaciones más cotizadas en el mercado mundial.</p>
VIABILIDAD PARA PYMES	<p>Esta es una solución para PYMES en el área de comunicaciones con un amplio nivel de probabilidad de éxito, esto debido a que SAS es una solución que lleva años trabajando en el sector además cuenta con el respaldo de ser una compañía multinacional que ofrece soluciones de acuerdo a las necesidades del servicio solicitado, para el caso de las PYMES en Colombia se acopla bastante bien , pues incluso casos como Bancolombia , Direc TV y hasta Cruz blanca utilizan tecnologías provenientes de SAS</p>

	Hadoop
VELOCIDAD	<p>Si bien hadoop permite administrar terabytes de datos, su velocidad de procesamiento se ve limitada a la de los discos que los almacenan. Especialmente cuando se trata de información tremendamente grande o requerida en tiempo real. A menos que se tenga la posibilidad de contar con arreglos de discos completamente SSD (Solid State Drives), algo que puede ser prohibitivo para muchas empresas debido a su costo, se estará a merced de las 7.200 revoluciones por minuto de los discos duros.</p>

VARIEDAD	<p>MapReduce está basado en tecnología de Google y se trabaja sobre discos , algunas variaciones como Spark corren en memoria. MapReduce es relativamente complejo de programar. Los desarrolladores pasaban mucho tiempo tratando de escribir codigos en MapReduce y resultaba en una enorme pérdida de tiempo sistemas SparkI3 cuenta con una API (Application Programming Interface) muy amigable para el desarrollador. Adicionalmente soporta ocho diferentes lenguajes de programación, incluyendo a Phyton, Java y R</p> <p>Crear algo realmente compatible. Spark puede funcionar sobre Amazon EC2 (Elastic Cloud 2), sobre Mesos de Apache y varios otros entornos cloud. Puede leer y escribir datos en una variedad de bases de datos, como PostgreSQL, Oracle, MySQL y con todos los formatos de archivos Hadoop.</p>
VOLUMEN	la capacidad de almacenamiento de las plataformas Hadoop depende directamente de el servicio data warehouse que se tenga o se contrate.
ARQUITECTURA	<p>Hadoop consiste básicamente en el Hadoop Common, que proporciona acceso a los sistemas de archivos soportados por Hadoop. El paquete de software The Hadoop Common contiene los archivos .jar y los scripts necesarios para hacer correr Hadoop. El paquete también proporciona código fuente, documentación, y una sección de contribución que incluye proyectos de la Comunidad Hadoop. Una funcionalidad clave es que para la programación efectiva de trabajo, cada sistema de archivos debe conocer y proporcionar su ubicación: el nombre del rack (más precisamente, del switch) donde está el nodo trabajador. Las aplicaciones Hadoop pueden usar esta información para ejecutar trabajo en el nodo donde están los datos y, en su defecto, en el mismo rack/switch, reduciendo así el tráfico de red troncal (backbone traffic). El sistema de archivos HDFS usa esto cuando replica datos, para intentar conservar copias diferentes de los datos en racks diferentes. El objetivo es reducir el impacto de un corte de energía de rack o de fallo de interruptor de modo que incluso si se producen estos eventos, los datos todavía puedan ser legibles.</p>
COSTO	El precio de las tecnologías de Hadoop depende directamente de 2 factores principales, el costo de arrendamiento de datos y el costo de los científicos de datos y programadores que estarán creando los scripts y manteniendo en operación el sistema

SECTOR ECONÓMICO	Debido a que hadoop es la tecnología mas genérica en el mercado en cuanto a Big Data se refiere, su campo de acción es muy amplio pues cualquier sector puede usar sus notables cualidades de procesamiento y el hecho que utilice todos los nodos de la red como puntos de procesamiento hace que para el sector estadístico sea un aliciente muy atractivo.
VIGENCIA EN EL MERCADO	Fue desarrollado originalmente para apoyar la distribución del proyecto de motor de búsqueda, denominado Nutch , su desarrollo ha sido notorio y a junio de 2010 la lista de sus sistemas operativos soportados incluía HDFS, Amazon S3, CloudStore, FTP, HTTP y HTTPS de solo lectura, su vigencia en el mercado hace que sea una de las herramientas usadas por las grandes corporación de internet como yahoo, quien en 2009 lanzo el hadoop web searchmap vigente como motor principal de búsqueda
VIABILIDAD PARA PYMES	Sin lugar a dudas hadoop es una de las herramientas de tratamiento de Big Data que más se acopla a las necesidades de las PYMES en Colombia en el especial las de comunicaciones, pues su basto nivel de procesamiento en tiempo real ofrece una característica que en las demás soluciones representa un costo elevadísimo, aunque cuenta con una dificultad y es encontrar el personal capacitado para programar las interfaces entre los datos y Hadoop pero en relación a lo que se quiere lograr es una opción a tener en cuenta por su nivel de simplicidad a la hora de la ejecución

	Hp vertica
VELOCIDAD	HP Vertica, diseñada pensando en la velocidad, la escalabilidad, la simplicidad y la apertura en su núcleo, ofrece velocidad ultrarrápida donde las consultas se ejecutan entre 500 y 1000 veces más rápidamente analizando la escala de Terabytes que almacena entre 10 y 30 veces más datos por servidor.
VARIEDAD	Todos los tipos de datos estándar básicos.(BIGINT, CHAR, DATE, DECIMAL, INTEGER, REAL, SMALLINT, VARCHAR)
VOLUMEN	La tecnología de Hp vertica ofrece 300 Terabytes de capacidad equivalentes a 300 billones de bytes.
ARQUITECTURA	<p>Arquitectura Híbrida In-Memory/On-Disk</p> <p>A diferencia de los sistemas tradicionales RDBMS que mantienen cerraduras pesadas en toda la base de datos mientras se cargan los datos, Vertica fue diseñado con un modelo transaccional de viaje único que garantiza consultas de concurrencia extremadamente alta, mientras se puede cargar nuevos datos en el sistemas de forma simultánea. Vertica a menudo es capaz de cargar datos hasta 10 veces más rápido que las soluciones tradicionales como resultado de este diseño. Además, Vertica fue construido con una arquitectura híbrida in-memory/on-disk para garantizar la disponibilidad en tiempo casi real de la información.</p>
COSTO	1200-1500 USD
SECTOR ECONOMICO	Cientos de clientes y un creciente ecosistema de socios confían en la Plataforma HP Vertica Analytics para monetizar el Big Data a altísimas velocidades y a escala masiva. Los clientes y los socios pueden obtener una certificación de Big Data y realizar un análisis comparativo del conocimiento de un usuario de la Plataforma HP Vertica Analytics con las credenciales ofrecidas a través del programa HP ExpertOne, las cuales pueden ser aplicadas para distintos sectores de la economía como manejo de volúmenes de datos en datacenter, o el estudio del mercado predictivo como ya se aplica en netflix u otros entes de reconocimiento mundial.

VIGENCIA EN EL MERCADO	En junio de 2013, HP anunció la plataforma de refugio para el análisis y la búsqueda de significado de las grandes ficha petabytes de información estructurada y no estructurada. HP Haven está impulsado por tres motores de análisis de la cartera de HP; Vertica, IDOL y distribuido R. Por ejemplo, la integración de IDOL y Vertica permite a los usuarios conectarse a una variedad de máquinas, negocios, fuentes de datos humanos y llevar a cabo tanto el análisis estándar y predictivo.
VIABILIDAD PARA PYMES	Los casos mas notables de empresas que hayan implementado esta tecnología, son de entidades de otros países donde los clientes empresariales, como Easynet Global Services y Gruppo Mediaset, han implementado la tecnología del servidor HP ProLiant como parte de una infraestructura convergente de HP para mejorar las eficiencias operativas y acelerar la entrega de aplicaciones dentro de sus entornos de TI.

	Data integrator Oracle
VELOCIDAD	Mediante el uso de Oracle Data Integration, las organizaciones han reducido sus costes de desarrollo en un 30% gracias a la mejora en la velocidad de datos de gestión en un 50%, y la reducción de los tiempos de ejecución de procesos de negocio por al menos el 70%.
VARIEDAD	Oracle maneja por su parte datos de la naturaleza de : DATE, FLOAT, NUMBER, TIMESTAMP y VARCHAR2,La precisión es necesaria para NUMBER a menos que establezca la propiedad de origen de datos, Si no especifica una precisión y no establece UseSQLToRetrieveSchema en TRUE, Campaign considera que puede almacenar los valores en un tipo de datos que conserve 15 dígitos de precisión. Esto puede ser problemático si los campos contienen valores cuya precisión sobrepasa los 15 dígitos, porque dicha precisión se perderá cuando el valor se incorpore a Campaign.
VOLUMEN	Data integrator por su parte permite una manipulación de datos cercana a los 1000 Terabytes, una capacidad que excede el triple de la capacidad de Hp vertica y que nos provee una aplicación mucho más amplia dado el alcance de su tecnología.

ARQUITECTURA	<p>Arquitectura Orientada a Servicios. E-LT podría definirse siguiendo el orden de las iniciales que lo denominan. Así se puede decir que consiste en la extracción, carga y transformación de datos, y se resume en los siguientes tres pasos:</p> <p>Primero: habrá que extraer y cargar los datos de manera “BULK” directamente a una Base de Datos o a unas tablas especialmente creadas para los datos de paso (conocidas también como staging). Esto supone que este medio servirá solo temporalmente, por lo que podrá ser limpiado en cada proceso de carga. Por ello se recomienda hacer transformaciones simples y limpieza básica de información.</p> <p>Segundo: cuando la información se halla contenida en staging habrá que proseguir con la elaboración del proceso de transformación de los datos, que posteriormente pasará a la base de datos del Data Warehouse. Esta transformación se hará con el lenguaje propio de la base de datos, por ejemplo T-SQL, PL/SQL.</p> <p>Tercero: una vez que se tienen los datos transformados en los procesos propios de la base de datos, se insertarían en el Data Warehouse. Terminada esta acción, se pueden limpiar los datos de paso, si se cree conveniente. De esta manera el proceso de transformación queda integrado en el motor de la Base de Datos.</p>
COSTO	1500 USD
SECTOR ECONÓMICO	<p>Uno de los campos mas potentes para la implementación de ODI yace en el almacenamiento en la nube, un campo que puede ser explotado por muchos entes de la economía, entidades como dropbox, facebook, icloud o google drive, para sacar provecho de las herramientas desde una perspectiva ODI, se necesita saber qué tipo de datos son compatibles. Puesto que queremos utilizar tipos de datos genéricos en lugar de los nativos, tenemos que tener un módulo de ingeniería inversa, que devuelve los tipos de datos genéricos y los hace manejables y aplicables a muchos conocimientos como se nombraron anteriormente.</p>
VIGENCIA EN EL MERCADO	<p>Oracle adquirió Sunopsis en octubre 2006 y volvió a ponerla en el mercado como Oracle Data Integrator (ODI). El objetivo de esta adquisición era ampliar la oferta de Oracle Fusion Middleware, el cual proveía amplio soporte a sistemas heterogéneos de origen y destino. Después de la adquisición Oracle continuó ofreciendo por separado ODI así como su anterior producto de ETL Oracle Warehouse Builder. En enero 2010 Oracle anunció su intención de fusionarlas en un solo producto (Oracle Data Integrator Enterprise Edition)</p>

VIABILIDAD PARA PYMES	<p>En el actual entorno empresarial cada vez más acelerado, las organizaciones deben utilizar las aplicaciones de software más especializadas y certeras, es necesario garantizar la coexistencia de estas aplicaciones en plataformas de hardware y sistemas heterogéneos, garantizar la capacidad de compartir datos entre aplicaciones y sistemas. ODI soporta las siguientes tecnologías y sistemas: Hypersonic SQL, IBM DB2/400, DB2 UDB, Informix, JD Edwards Enterprise One, JMS, Oracle Database, Oracle Data Quality for Data Integrator, SAS, Sybase ASE, Sybase IQ , Teradata, Web Services (Metodo GET), LEmail Y SFTP, entre otras, muchas tecnologías de las cuales se han visto implementadas y han sido descritas con anterioridad como es el caso de SAS o Teradata.</p>
------------------------------	---

	IBM Solutions/C.Datos
VELOCIDAD	<p>IBM anunció que logró nuevo avance tecnológico que ayudará a mejorar las velocidades de Internet a 200 - 400 Gigabits por segundo con un consumo de energía extremadamente bajo. El aumento de la velocidad se basa en un dispositivo que puede usarse para mejorar la transferencia de <i>Big Data</i> (Grandes Volúmenes de Información) entre las nubes y los centros de datos cuatro veces más rápido que la tecnología actual. A esta velocidad de 160 Gigabytes, el equivalente de una película de ultra alta definición de 4K de dos horas, ó 40,000 canciones; podrían descargarse en sólo unos cuantos segundos. El dispositivo se presentó en la Conferencia Internacional de Circuitos de Estado Sólido (ISSCC) en San Francisco. Cabe resaltar que esta tecnología es por hoy un prototipo.</p>
VARIEDAD	<p>Si bien podemos hablar de variables de tipo BIGINT, CHAR, DATE, DECIMAL, INTEGER, REAL, SMALLINT, VARCHAR, entre otros tano el big data de IBM como el de HP VERTICA u ODI están listos para manipular tipos de datos de fuentes diversas que traducidos en un lenguaje más coloquial harían referencia a los datos ligados a:</p>
VOLUMEN	<p>Web and Social Media: ,Machine-to-Machine (M2M, Big Transaction Data, Biometrics, Human Generated, tipos de datos cuya descripción ya ha sido detallada en el escrito y que describen los alcances del big data actual</p>

ARQUITECTURA	Grandes volúmenes de datos se pueden almacenar, adquieren, procesan y analizan en muchos aspectos. Cada fuente de datos grande tiene características diferentes, incluyendo la frecuencia, el volumen, la velocidad, el tipo y la veracidad de los datos. Cuando los grandes datos se procesan y almacenan, dimensiones adicionales entran en juego, tales como gobernabilidad, seguridad y políticas. La elección de una arquitectura y la construcción de una solución de grandes volúmenes de datos apropiada es difícil porque hay muchos factores tienen que ser considerados. Esta serie de "arquitectura de datos grande y patrones" presenta un enfoque estructurado y basado en patrones para simplificar la tarea de definir una arquitectura global de datos grande. Debido a que es importante evaluar si un escenario de negocios es un problema de grandes datos, incluimos sugerencias para ayudarle a determinar qué problemas de negocios son buenos candidatos para las soluciones de datos grandes.
COSTO	0 USD
SECTOR ECONOMICO	Tiene múltiples aplicaciones en la industria, por ejemplo en la multimedia permite la ubicación del usuario móvil y focalización de precisión, la identificación de datos generados por máquina o incluso aplicado a el servicio de citas en línea donde se realiza un análisis sofisticado para medir la compatibilidad entre miembros individuales y obtener un mejor resultado. Pero no solo eso, IBM provee la facultad de aplicar el Big Data a servicios públicos, por ejemplo en la industria del cable, los datos grandes pueden ser utilizados para analizar datos de decodificadores diariamente por grandes operadores de cable tales como Time Warner, y Cox Communications. Estos datos pueden ser aprovechados para ajustar la actividad de publicidad o promocional, o incluso en entidades financieras evitando el Fraude de pago con tarjeta de crédito y débito, el Fraude de depósito de cuenta o el Fraude técnico y deuda incobrable, entre otros.
VIGENCIA EN EL MERCADO	Vigente desde 1993
VIABILIDAD PARA PYMES	<p>Las empresas demandan acceso a grandes volúmenes de es por eso que las soluciones Big Data de IBM, junto con las soluciones Hadoop de empresa, permiten a los usuarios almacenar, gestionar y analizar datos de muy diversas fuentes, además de dar acceso a los datos a analistas de negocio, científicos de datos y usuarios de TI a través de sus múltiples modelos aplicables a los negocios:</p> <p>IBM BigInsights for Apache Hadoop Permite a las empresas aprovechar grandes volúmenes de datos variados y analizarlos de forma rápida, sencilla y rentable.</p> <p>IBM BigInsights on Cloud Proporciona Hadoop como servicio a través de la infraestructura cloud de IBM SoftLayer</p> <p>IBM InfoSphere Streams Permite a las empresas capturar y analizar datos en movimiento, así como extraer rápidamente información de valor para aplicaciones críticas del Internet de las cosas.</p>

CAPITULO 7

MODELO DE IMPLEMENTACIÓN DE PROYECTOS DE BIG DATA ORIENTADO A PYME DE COMUNICACIONES

El análisis para la implementación de Big Data en cualquiera de las áreas mencionadas o en cualquier entidad PYME requiere establecer un modelo de análisis donde se tenga en cuenta las particularidades que pueden repercutir de manera positiva o negativa en la adaptación del sistema a una actividad económica en especial. El modelo de análisis permite describir un proceso a partir de texto, lo cual hace posible examinar los requisitos para la implementación de la tecnología, aumentando la comprensión frente al tema.

Es bien sabido que la implementación del Big Data en esta investigación busca dar solución al inconveniente que presentan las PYMES en cuanto al manejo de los datos y a la explotación de los factores de variedad, velocidad y volumen que requiere un gran flujo de datos en un sistema en constante evolución, razón por la cual se genera un análisis conveniente para dar solución a los alicientes de las actuales entidades generadas por una cantidad de datos creciente y que sin las medidas tecnológicas adecuadas parecería irresoluble.

7.1 MODELO DESCRIPTIVO PARA ESTABLECER LA VIABILIDAD DE UN PROYECTO BIG DATA

En un informe reciente del Centro IBM para los asuntos de gobierno, Kevin C. Desouza de la Universidad Estatal de Arizona⁴⁰, estableció un modelamiento para la implementación de Big Data en el sector público que consta de tres etapas, cada una de ellas con pasos a seguir para que la ejecución de Big Data sea una realidad. Este modelo o plan de acción sirvió como soporte para establecer el modelo de implementación de Big Data para PYMES en Colombia.

ETAPA 1: PLANIFICACIÓN DEL PROYECTO BIG DATA:

Sin duda alguna uno de los factores más determinantes en el éxito o el fracaso de un proyecto es la planificación del mismo, en el caso de la implementación de Big data no resulta ser diferente, puesto que determinar si su inclusión en una organización requiere grandes cambios de infraestructura, software o personal entre otros, esta etapa comprende la conceptualización de la idea lo cual es primordial para la materialización del proyecto.

PASO 1: INVESTIGAR EL CAMPO DE ACCIÓN:

Es necesario determinar qué cantidad de datos maneja la organización, cuales son los riesgos y desafíos inmersos en este emprendimiento, esto implica reconocer si la cantidad de datos que genera la organización es exorbitante o elevada, del nivel de los terabytes o superior, de tal manera que justifique los costos que necesita el Big data o si de lo contrario puede buscarse una solución más económica pero igual de eficiente debido a que la cantidad de datos no requiere mayor inversión.

PASO 2: CREAR UNA COALICIÓN PARA APOYAR EL PROYECTO.

Es conveniente generar una alianza en donde se difunda la idea de manera más fluida, hablar con colegas, instituciones académicas o centros de información pueden ampliar el concepto acerca de la tecnología y sacar a luz soluciones más prácticas a la problemática.

PASO 3: DEFINIR METAS

Se traduce en desarrollar la idea entorno a los implicados de manera benéfica, esto quiere decir que mediante el proyecto se buscará beneficiar a la comunidad directamente implicada en el desarrollo del proyecto de manera que se atraiga su atención y se dé lugar a un pensamiento crítico que beneficie la iniciativa.

PASO 4: ALINEACIÓN ESTRATEGICA DE LOS PROYECTOS DE BIG DATA:

Cuando se implementa una nueva tendencia se teme que los esfuerzos de la organización se centralicen en esta nueva idea y que los proyectos que ya estaban en curso pierdan su rumbo, una forma de lidiar con estas inseguridades es encaminando la implementación de la tendencia Big Data a los antiguos proyectos, esto es en otras palabras, entretener la tendencia con los proyectos de la organización y en la medida de lo posible con un alto ejecutivo dispuesto a defender la idea en los momentos de duda, como lo señala Desouza.

PASO 5: PRIVACIDAD Y SEGURIDAD:

Los CIO u oficiales en jefatura de sistemas deben estar preparados para asumir responsabilidades en cuanto a la seguridad que ofrece Big Data, la privacidad y las consideraciones éticas alrededor de la recopilación, integración, análisis y difusión de datos, deben ser discutidos abiertamente y con sinceridad a medida que el proyecto suma importancia, esto le dará mayor credibilidad y mitigará futuras trabas en el proyecto.

PASO 6: INDICADORES DE RENDIMIENTO:

Desarrollar indicadores de rendimiento es una herramienta útil para cualquier iniciativa, los indicadores de rendimiento enfocados en eficiencia y resultados permiten determinar la viabilidad del proyecto, las metas alcanzadas y el margen de rentabilidad que tiene esto para la empresa.

ETAPA 2: EJECUCIÓN DEL PROYECTO BIG DATA:

Una vez planificado el rumbo que se desea que tome la iniciativa y trazadas las metas a corto, mediano y largo plazo es momento de poner las ideas en marcha para materializar el concepto.

PASO 7: REQUISITOS BIG DATA:

Existen muchos factores a tener en cuenta los cuales nos permiten determinar si la implementación de Big Data en una organización es requerida como ya fue descrito en la planificación del proyecto, estos datos son de carácter diverso e implican la cantidad de datos, la disposición de expertos y el conocimiento del negocio para sacar provecho, entre otras fuentes de información. Es por eso que una vez definido el rumbo del proyecto se enuncian algunos de los principales requisitos para implementar Big data en el siguiente diagrama:



PASO 8: CALIBRACIÓN DE LA INICIATIVA BIG DATA:

Los CIO deben ponerse constantemente al frente de la iniciativa, los problemas y a su vez fomentar soluciones creativas a los inconvenientes que se presenten con el pasar del tiempo. Las comunicaciones apropiadas y oportunas, junto con otras intervenciones, pueden ayudar a abordar las cuestiones y cortar de raíz los problemas potenciales

PASO 9: MANEJAR LA CORRUPCIÓN:

Dado que los proyectos de Big Data se desarrollan para entidades con grandes volúmenes de información y de economía influyente es probable que a medida que se desarrolla el proyecto y se obtienen resultados los integrantes de la organización quieran sacar provecho de la tecnología utilizándola para otros fines, es por eso que es necesario analizar incluso los factores humanos para el éxito de la inclusión Big Data en una PYME, es por ello que resulta necesario que los CIO mantengan un ojo vigilante sobre la corrupción del alcance y tengan en claro los límites del esfuerzo actual y cómo se tomarán las revisiones y adiciones futuras

PASO 10: EL ENFOQUE EN LOS DATOS:

Las tecnologías de Big Data están evolucionando a un ritmo excepcional. Sin embargo, puede que el proyecto no necesite toda la nueva tecnología. Tal vez puede cambiar la finalidad de los activos tecnológicos existentes. Mantenerse enfocado en la gestión de los datos es lo primordial. Con una visión clara de los problemas de gestión de datos a partir de un punto de vista organizativo y de la política, elegir la tecnología adecuada debería ser relativamente fácil. Tener el enfoque claro, aprovechar los datos y gestionarlos de buena manera son algunos de los puntos que ayudaran a mantener el rumbo del proyecto.

PASO 11: IMPLEMENTACIÓN DEL MODELO

Una vez realizado todo el estudio previo incluido en la planificación del proyecto y tenidos en cuenta todos los factores influyentes en la etapa de ejecución se describe de manera detallada los pasos a seguir para implementar Big data en una PYME, en el modelo desarrollado y descrito posteriormente.

ETAPA 3: POST-IMPLEMENTACIÓN DE SU PROYECTO DE BIG DATA

Una vez que se ha logrado estructurar un sistema funcional de Big Data el trabajo aún no está completo, es necesario analizar constantemente la actividad para verificar que todo marche de la manera adecuada.

PASO 12: REALIZAR UN ANÁLISIS POST-MORTEM DEL PROYECTO BIG DATA

Es buena idea documentar todo el proyecto, incluyendo las lecciones aprendidas de todas las etapas, para retener el conocimiento institucional que obtuvo como resultado. Esta información también puede ser compartida con la sociedad o los neófitos en el tema con el fin de afianzar un camino a seguir para una implementación exitosa.

“Un elemento importante de la realización de un análisis port-mortem es que no se debe utilizar para la evaluación o para apuntar cifras en personas o acontecimientos”, señala Desouza. “A menos que las personas están protegidas sobre compartir sus verdaderas experiencias y episodios de aprendizaje, el ejercicio post-mortem no será de ningún valor

PASO 13: IDENTIFICAR SU PRÓXIMO PROYECTO DE BIG DATA

Los resultados de un proyecto de Big Data y las lecciones aprendidas deben ayudar a identificar oportunidades para nuevos proyectos de grandes datos. La toma de decisiones en tiempo real y el comportamiento predictivo son solo algunos de los beneficios que puede otorgar el Big Data y de los que se puede sacar provecho para múltiples funciones.

7.2 MODELO PROPUESTO PARA IMPLEMENTACIÓN DE PROYECTOS DE BIG DATA ORIENTADO A PYME DE COMUNICACIONES

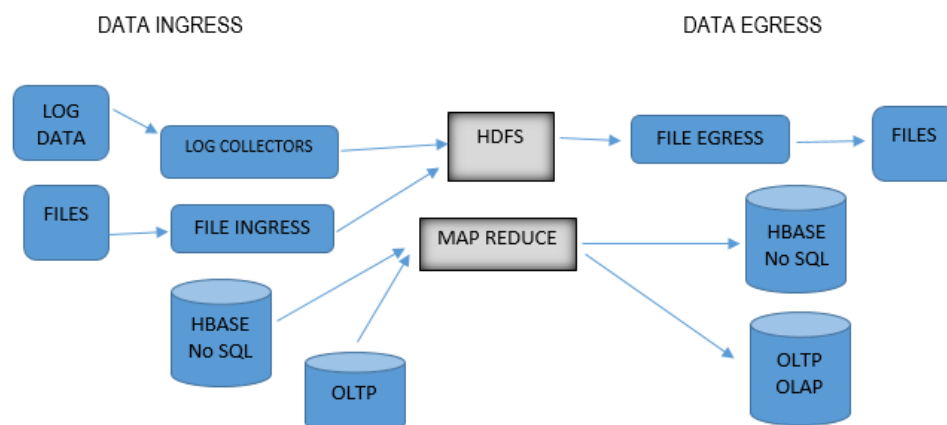
Para las Pymes en Colombia el factor más influyente es la inversión de capital en la adquisición de nuevas tecnologías, como se aclarara en apartados posteriores existen en el mercado varias opciones a la hora de implementar Big Data, es por eso que son necesarias pautas para el conocimiento de cómo hacer la transición entre la forma tradicional de análisis y las ofrecidas por Big Data y cuál será el modelo a seguir. El siguiente modelo descriptivo propuesto, toma como contraste el modelo de Kevin C. Desouza mencionado al detalle en el apartado anterior, para esta monografía se tomaron en cuenta las necesidades y especificaciones que podrá presentar una PYME de comunicaciones en Colombia y adicionándole los conceptos generales que deben ser cubiertos desde el punto de vista de meta data, la analítica de datos, almacenamiento , seguridad y necesidades de usuario final, todas hacen parte del conjunto global que se define como Big Data.

7.2.1 INDEXACIÓN UNIVERSAL EN TIEMPO REAL DE CUALESQUIERA DATOS DE MÁQUINA

Este es el núcleo de Big Data conforme la mayoría de las personas conciben; usualmente es igualado con el proyecto de código abierto Hadoop. No hay que confundir la indexación en Hadoop con un índice en la base de datos relacional: Un índice Hadoop es un índice de archivos. De esta manera, Hadoop puede ingerir muchos tipos diferentes de datos y convertirlos en bases de datos JSON para plataformas como MongoDB por citar un ejemplo.

Las compañías ya pueden ser inundadas con alimentaciones de movimiento de ID de radio- frecuencia (RFID), clic en sitios web, y otros datos que pueden ser estructurados si el personal de TI gasta algún tiempo para convertirlos en datos estructurados y los coloca en una base de datos relacional. Eso puede valer la inversión si la empresa sabe cómo los datos serán usados y cómo serán consultados y accedidos en el futuro.

Sin la necesidad de que la empresa tenga que conocer el futuro potencial de los datos, Hadoop proporciona una salida. Al tomar los datos tal y como vienen, Big Data genera una postergación a la definición de datos, para cuando se realiza el análisis. Hadoop por su parte distribuye los datos a lo largo de muchos servidores y da seguimiento a las ubicaciones sin limitar el uso futuro.



7.2.2 BÚSQUEDA LIBRE Y ANÁLISIS DE HISTORIAL EN TIEMPO REAL

El almacenamiento de los datos es solo parte de la manera de tener éxito. Pues la información necesita ser relativamente fácil de recuperar. La manera más rápida de hacer eso es proporcionando una posibilidad de búsqueda que sea rápida (como en la implementación, no tiempo de respuesta). Abra que buscar un juego de herramientas que permite búsquedas de texto de datos no estructurados. Apache

Lucen⁴⁰ es una herramienta común que proporciona indexación de texto en un entorno de Big Data.

En adición el tener una respuesta en el monitor es decir una interfaz final con el usuario brinda al personal un nivel de confiabilidad al saber que todo está siendo almacenado de la manera correcta y puede ser accedido. El pasó administrativo para esto es indexar el contenido de los datos almacenados en los nodos distribuidos. Las consultas de búsqueda posteriormente acceden a los índices en los nodos distribuidos en paralelo para proporcionar una respuesta más rápida.

7.2.3 DESCUBRIMIENTO DE CONOCIMIENTO AUTOMATIZADO DE LOS DATOS

De la misma manera como puede ser ineficiente trasladar todos los datos semi-estructurados hacia una base de datos relacional, la realización de búsquedas manuales e informes manuales es ineficiente para el análisis. Las herramientas de minería de datos y análisis predictivos están rápidamente convirtiéndose para poder usarse en Big Data como una fuente de datos para análisis y para ser una base de datos para monitorear continuamente el cambio. Todas las herramientas de minería de datos siguen este proceso. Alguien determina el propósito del análisis, observa los datos, y posteriormente desarrolla modelos estadísticos que proporcionan conocimiento o realizan predicciones. Esos modelos estadísticos posteriormente necesitan ser desplegados en el entorno de Big Data para realizar evaluaciones continuas. Esta porción debe ser automatizada.

7.2.4 MONITOREO DE DATOS Y ALERTAS EN TIEMPO REAL

La empresa deberá buscar una herramienta para monitorear los datos en Big Data. Existen herramientas que crean consultas que son continuamente procesadas, buscando criterios a ser cumplidos. Asumiendo que la mayoría de los datos entrantes son no estructurados y no están destinados a una base de datos relacional, el monitoreo en tiempo real es probablemente el área donde un elemento de datos es inspeccionado más estrechamente.

Por ejemplo, la empresa puede establecer una alerta para cuando el chip RFID⁴¹ en un artículo de comida congelada sea almacenado en un área no congelada. Esa alerta puede ir directamente a dispositivos móviles que son usados en el almacén, previniendo que la comida se arruine. Los movimientos de los clientes en una tienda también pueden ser monitoreados y la publicidad puede ser dirigida hacia el cliente preciso que esté parado frente a un artículo específico en monitores posicionados estratégicamente.

7.2.5 DEBE PROPORCIONAR INFORMES Y ANÁLISIS AD HOC

En relación al descubrimiento de conocimiento y minería de datos automatizada, los analistas necesitan acceso para recuperar y resumir la información en el entorno de nube de Big Data para el caso del cloud. La lista de proveedores cuyas herramientas funcionan para reportar desde Big Data para crecer más cada día. Algunas de las herramientas usan Apache Hive⁴² y Hive Query Language⁴³. Las sentencias HQL son similares a las sentencias de Structured Query Language y muchas de las herramientas que proporcionan estilos familiares de informes a partir de Big Data usan la interfaz HQL y Hive para ejecutar las consultas a través de MapReduce.

7.2.6 LA EMPRESA DEBE TENER LA CAPACIDAD DE CREAR RÁPIDAMENTE PANELES DE INSTRUMENTOS Y VISUALIZACIONES PERSONALIZADAS

Como en la evolución de un proyecto de inteligencia de negocios tradicional, cuando las personas pueden consultar Big Data y producir informes, ellos desean automatizar esa función y crear un panel de instrumentos para visualizaciones repetitivas con fotografías amenas y de fácil entendimiento, la mayoría de las herramientas tienen cierta habilidad para crear visualizaciones tipo dashboard desde sus sentencias de consulta.

Es un poco prematuro pensar en la implementación de Big Data si la empresa no puede cumplir con una predicción, con base en la historia de la inteligencia de negocios, pues los paneles de instrumentos son la clave para la integración de elementos en Big Data. Y de acuerdo a la historia de la inteligencia de negocios, el contar con buenos paneles de instrumentos para Big Data será importante para obtener y mantener el soporte ejecutivo.

7.2.7 SEGURIDAD Y CONTROLES DE ACCESO

Cuando los datos no estructurados están en un mundo de datos relacionales, la complejidad de acceso a los datos puede impedir que todos obtengan los datos. Las herramientas de informes comunes no funcionarán. El traslado hacia Big Data representa un paso activo hacia hacer que lo complejo sea accedido más fácilmente. Desafortunadamente, las mismas configuraciones de seguridad usualmente no se traducen de los sistemas relacionales existentes hacia los de Big Data. El tener buena seguridad se volverá más importante mientras más Big data sea usado. Conforme la compañía desarrolle más análisis usando los datos de Big Data, a los resultados se les deberá dar seguridad, particularmente a los informes y paneles de instrumentos, de

manera similar como se brinda seguridad a los informes de los sistemas relacionales actuales.

Al iniciar con Big Data basados en la nube, se debe estar consciente de la necesidad de aplicar seguridad en algún punto, particularmente al entorno de informes y de paneles de instrumentos.

7.2.8 SE DEBE DAR SOPORTE A LA IMPLEMENTACIÓN FLEXIBLE Y DE MULTI-INQUILINOS

Muchas personas tienen recelo para colocar los datos críticos en un entorno de nube. El punto es que la nube proporciona el bajo costo y la rápida implementación que se necesitan para iniciar proyectos de Big Data. Precisamente debido a que el proveedor de nube colocará los datos en una arquitectura donde los recursos de hardware son compartidos, el costo es dramáticamente menor.

De una manera ideal cada empresa quisiera tener sus datos únicamente bajo el cuidado de sus propios servidores y que algún ente externo realizara la gestión y la configuración. Sin embargo, eso simplemente no es un modelo de negocios rentable cuando las necesidades de Big Data son intermitentes. El resultado es mayor costo debido a que las compañías pagarían por mucho tiempo inactivo, especialmente durante los primeros proyectos, cuando los analistas están explorando, probando y aprendiendo Big Data.

7.2.9 INTEGRACIÓN EXTENSIBLE VÍA APIS

El Big Data fue diseñado para ser accedido por aplicaciones de producto. Los métodos comunes de acceso usan interfaces de programación de aplicación (APIs) RESTful⁴⁴. Estas están disponibles para cada aplicación en el entorno de Big Data para control administrativo, almacenamiento de datos, e informes de datos. Debido a que todos estos contenidos fundacionales de Big Data son de código abierto, estas APIs están bien documentadas y están abiertamente disponibles para su uso. Los proveedores de Big Data basados en la nube permitirán el acceso a todas las APIs actuales y futuras bajo la seguridad apropiada

CAPITULO 8

COSTOS ESTIMADOS PARA PYMES DE COMUNICACIONES

8.1 CONSIDERACIONES PARA IMPLEMENTAR BIG DATA (COSTOS)

Es necesario entender en primera instancia que si bien el Big Data es una alternativa eficiente para el manejo de grandes volúmenes de datos, este no implica que en todos los casos sea la mejor opción, esto se fundamenta en el hecho de que en algunos casos las empresas por estar al corriente de las tendencias pueden considerar implementar la tecnología en sus sistemas aun cuando el sistema para el manejo de datos que estén utilizando sea eficiente y no representa un mayor costo, dicho de otra manera el implementar esta tecnología no se tiene como propósito realizar una inversión injustificada o precipitada, por lo cual es conveniente realizar un análisis previo para evidenciar si el sistema recibirá los beneficios ante la implementación de este o si por el contrario representa costos muy elevados para trabajos que otros medios pueden manejar con una eficacia equivalente⁴⁵. Además de lo dicho anteriormente puede hacerse notar el beneficio que otorga el desarrollo del Big Data al ser estructurado sobre software libre, puesto que se tiene soporte para el desarrollo de las tecnologías basados en este, una vez expuesta la necesidad de realizar un análisis previo en las entidades que desean implementar este servicio y las facilidades que pueden preverse gracias al software libre es posible establecer los factores influyentes en el análisis de viabilidad y costos del Big Data.

8.1.1 CULTURA ORGANIZACIONAL

Aunque no es uno de los parámetros que se aprecien de manera intrínseca a los costos del proyecto es necesario tener en cuenta que la falta de conocimientos ante el tema representa una falencia que requiere ser reparada, traduciéndose en capacitaciones, estudios y preparación, lo cual implica inversiones para adaptar el personal a la nueva tendencia.

8.1.2 INTEGRACIÓN DE LAS ÁREAS DE LA ENTIDAD

Si bien las diversas áreas de una entidad ya manejan cierta relación profesional con las demás por el beneficio del negocio para implementar el Big Data de manera exitosa es necesario que cualquier agente que necesite la información en un momento dado pueda

acceder a ella de manera clara y transparente, lo cual implica sostener la información en bases de datos de una manera estructurada para evitar cualquier percance o confusión con los datos, es evidente que para que las áreas de una entidad se comuniquen y tengan acceso de manera equánime requiere un cambio en la infraestructura o al menos en la gestión de los datos, lo cual representa también una inversión por parte del empresario.

8.1.3 INFRA ESTRUCTURA Y NODOS

Este representa uno de los puntos más evidentes y determinantes en el costo de una iniciativa Big Data dado que la potencia del sistema⁴⁶ está relacionado con la potencia de cálculo de los nodos en el sistema, no obstante gracias al avance de la tecnología cada vez se consiguen equipos con mejores especificaciones y a más bajo costo lo cual hace que promover su implantación sea cada vez más rentable para así lograr obtener un sistema con éxito.

Por ello para obtener un sistema eficaz y más económico las entidades deberán contar con metas específicas y métricas adecuadas para el cálculo de la capacidad del sistema y cuidándose de no realizar inversiones innecesarias de modo que al trazar un plan ROI⁴⁷ este se vea reflejado de manera positiva sobre la organización.

Si la entidad cuenta con una infraestructura previa el costo de la innovación podría verse ampliamente reducido, de modo tal que las instalaciones previas de la organización también son influyentes en el costo de la implementación.

8.2 COSTOS ESTIMADOS PARA PYMES DE COMUNICACIONES

Si bien es posible estimar el costo que implica establecer una estructura bien ordenada para el manejo y procesamiento de grandes volúmenes de datos, es preciso detallar que el costo varía por factores que van desde la evidente infraestructura hasta la preparación del personal para la operación de la información. Sin embargo no son estos los únicos factores que influyen en la estimación de un costo prudente para una instalación de este tipo, en estos análisis se deben incorporar además las consideraciones que establecen la existencia de una infraestructura previa al desarrollo, es decir, si la empresa que desea implementar Big data ya posee un centro de manejo de datos menos avanzado o ha implementado previamente de manera superficial la tecnología refiere una disminución notable en los costos. Para el caso puntual vamos a establecer los costos para una empresa con un volumen de datos considerable con un equipamiento de diez a quince racks, los costos en el diseño, implementación y operación implican el estudio para minimizar los costos en cuanto a

la infraestructura y diseño.

Los campos a considerar en el análisis de costos son el número de racks y sus accesorios, el cableado estructurado que requiera la obra, la refrigeración para los equipos, la acometida eléctrica, los sistemas de supresión de incendios (opcional) y los sistemas de seguridad, esto dentro de la parte de la infraestructura, en cuanto para el manejo de los equipos se ha de tener en cuenta el personal, la formación de entes idóneos, el mantenimiento y respectivas reparaciones que puedan implicar los sistemas en un futuro.

En primera estancia se considera un espacio de 35 m² destinado a distribuir 14 bastidores o racks con el fin de albergar todos los equipamientos electrónicos y de comunicaciones requeridos para establecer el centro de manipulación de los datos, el espacio es conveniente considerando que los bastidores tienen medidas estandarizadas para lograr mayor compatibilidad entre los soportes y los equipamientos de diversos fabricantes, considerando la prevención de fallos energéticos y de climatización e implementando sistemas de refrigeración para el sistema una infraestructura como la descrita costaría 423.000 € según el Datacenter Capital Cost, lo cual se traduce en un costo por unidad de 30215 € aproximadamente, mientras que el arrendamiento financiero a un plazo de cuatro años de los equipos requeridos implicaría un costo que ronda los 475.000 €, sin contar con el costo requerido en el alquiler del espacio, sala o salón destinado para la disposición de los racks, pero estableciendo el centro de datos en completa operación para el manejo de datos. Sin embargo esta no es la única opción, existen algunas empresas que proveen el servicio de datacenter aprovisionado de servidores, equipos de comunicación, redes, ups, sistemas operativos, sistemas de respaldo y demás. De igual modo existe la opción de acudir a servicios de outsourcing para el almacenamiento, respaldo y soporte técnico por entidades externas dedicadas al servicio.

Solucion O prestador de Servicio	Capacidad Estimada	Precio (USD)
Data integrator de Oracle	1000 Terabytes	1500
HP Vertica	300 Terabytes	1200-1500
SAS Latinoamerica	2,5-3 Peta bytes	800-1000
Senseta	2-3 Terabytes	800-1200
Hadoop/C.Datos	Peta bytes	0
IBM Solutions/C.Datos	Peta bytes	0

Datacenter	150 Petabytes	1100000
------------	---------------	---------

Tabla N°1 Comparativa Costo - Capacidad



Prestador de servicio de Cloud Computing, Todas Basadas en Hadoop.



Soluciones de Tipo FrameWork, Son imprescindibles los C.Datos(Científicos de datos) Encargados de el análisis de la información y la creación de interfaces, en adición se debe contar con la infraestructura de servidores y clúster necesaria.



Datacenter y montaje físico de equipos.

A continuación se presenta un esquema de costos generado especialmente para pymes de comunicaciones en Colombia:

8.3 PRELIMINARES

COSTOS DE UNA IMPLEMENTACION DE BIG DATA PARA PYME EN COLOMBIA.				
ITEM	DESCRIPCION	UNIDADES	TIEMPO DE EJECUCION	COSTO APROXIMADO EN PESOS
	INSTALACIONES ELECTRICAS Y ELECTRONICAS NO REGULADAS			
1,00	PRELIMINARES			

1,01	OBRAS PRELIMINARES. (SUMINISTRO DE MANO DE OBRA PARA EL LEVANTAMIENTO Y RETIRO DE CIRCUITOS ELÉCTRICOS EXISTENTES Y EQUIPOS ASOCIADOS A ESTOS DESDE TABLERO DE PISO. DESCABLEADO DE PUNTOS ELÉCTRICOS PROVISIONALES E INHABILITADOS, COMO SON LAS TOMAS DE MURO Y PISO EXISTENTES ASI COMO LAS LINEAS TELEFONICAS NO EXISTENTES, DESCABLEADO DE PROVISIONALES SE DEBEN DESCONECTAR Y DESCABLEAR Y DESMONTAR TUBERIA DESDE EL PUNTO ELÉCTRICO ANTERIOR A LA SALIDA A INHABILITAR QUE SE VEA AFECTADA POR LA OBRA CIVIL.)	0	15 - 30 días	\$ 769.726,15
1,02	OBRAS PRELIMINARES (CONSIDERA UN ESPACIO DE 35 M2 DESTINADO A DISTRIBUIR 7 BASTIDORES O RACKS CON EL FIN DE ALBERGAR TODOS LOS EQUIPAMIENTOS ELECTRONICOS Y DE COMUNICACIONES REQUERIDOS PARA ESTABLECER EL CENTRO DE MANIPULACIÓN DE LOS DATOS, EL ESPACIO ES CONVENIENTE CONSIDERANDO QUE LOS BASTIDORES TIENEN MEDIDAS ESTANDARIZADAS PARA LOGRAR MAYOR COMPATIBILIDAD ENTRE LOS SOPORTES Y LOS EQUIPAMIENTOS DE DIVERSOS FABRICANTES, CONSIDERANDO LA PREVENCIÓN DE FALLOS ENERGÉTICOS Y DE CLIMATIZACIÓN E IMPLEMENTANDO SISTEMAS DE REFRIGERACIÓN PARA EL SISTEMA)	0	1-3 días	--
1,03	OBRAS PRELIMINARES (TOMAR EN CUENTA ASESORIA DE INGENIEROS EN TELECOMUNICACIONES Y/O SISTEMAS PARA ORIENTACION ADECUADA DE NUMERO DE RACKS)	0	1-5 días	--

Tabla N°2 Costos Preliminares

8.4 TABLEROS ACOMETIDAS Y CABLEADO ESTRUCTURADO

2,00	TABLEROS, ACOMETIDAS Y CABLEADO ESTRUCTURADO			
2,01	INSTALACIONES ELÉCTRICAS (SUMINISTRO E INSTALACIÓN DE TABLERO PRINCIPAL DE 4 CTOS TRIFASICO CON ESPACIO PARA TOTALIZADOR CON PUERTA, CHAPA Y CERRADURA EL CUAL SE UBICARÁ EN EL EL PISO , INCLUIRÁ BARRAJE DE NEUTRO Y TIERRA INDEPENDIENTES.. EL TABLERO DEBE QUEDAR EN BUENAS CONDICIONES FÍSICAS Y ELÉCTRICAS DE ACUERDO A LA DISPOSICIONES DEL RETIE. EL TABLERO SERA ORGANIZADO INTERNAMENTE, BALANCEAR CARGAS, DEBE PRESENTARSE CERTIFICADO DE CONFORMIDAD DEL PRODUCTO. E INSTALAR EL CUADRO DE CARGAS Y DEJAR COMPLETAMENTE IDENTIFICADO EL TABLERO. SE DEBE INCLUIR REGATAS, RESANES TERMINADO Y PINTURA).	0	5-10 días	\$ 1.008.665,45
2,02	INSTALACIONES ELECTRICAS (SUMINISTRO E INSTALACION DE INTERRUPTOR INDUSTRIAL DE 3X80A 25 KA 240V PARA UBICAR EN EL TABLERO DEDISTRIBUCION GENERAL DE CIRCUITOS)	1	1-2 días	\$ 524.238,46
2,03	INSTALACIONES ELÉCTRICAS (.SUMINISTRO E INSTALACIÓN DE CABLE NO.3 NO 4+1 NO. 2+1NO. 6T THHN-THWN 600V EN DUCTERIA DE 1 1/2"	Según espacio de 35 m2	1-3 días	\$ 54.499,39
2,04	CABLEADO ESTRUCTURADO (HORIZONTAL SEGÚN NORMA EIA/TIA 568A , CABLES DE PAR TRENZADO BLINDADO (STP) DE 150 OHMIOS Y CUATRO PARES)	Precio por metraje y cantidad de conexiones	1-5 días	\$ 1.200.000,00
2,05	CABLEADO ESTRUCTURADO (BACKBONE O VERTICAL SEGÚN LA NORMA EIA/TIA 568 SE USARA CABLE UTP, EL MISMO SERÁ DE CATEGORÍA 5E, 6 O 6A Y SE DISPONDRÁ UN NÚMERO DE CABLES DESDE CADA GABINETE AL GABINETE SELECCIONADO COMO CENTRO DE ESTRELLA.)	Precio por metraje y cantidad de conexiones	1-5 días	\$ 1.450.000,00
2,06	INSTALACIONES ELÉCTRICAS (SUMINISTRO DE MATERIAL Y MANO DE OBRA DE INSTALACION PARA INTERRUPTOR TERMOMAGNETICO ENCHUFABLE DE 3X50A(SE VERIFICARA CON EL SUPERVISOR LOS INTERRUPTORES A SUMINISTRAR).	1	1 día	\$ 49.107,14

Tabla N°3 Costos Acometida y Cableado

8.5 ILUMINACIÓN Y TOMAS

3,00	ILUMINACION Y TOMAS			
3,01	INSTALACIONES ELÉCTRICAS. (SUMINISTRO DE MATERIALES EN GENERAL Y EJECUCIÓN DE RECABLEADO DE SALIDAS ELÉCTRICOS PARA ILUMINACIÓN, SE INCLUYE: DESCONEXIÓN , RETIRO Y CABLEADO ELÉCTRICO EXISTENTE DE LA ACTUAL ILUMINACIÓN, ADECUACION DE LAS CAJAS DE SALIDAS Y TERMINALES , CON SU CORRESPONDIENTE ATERRIZAJE O LÍNEA A TIERRA EN ALAMBRE DESNUDO Nº 12, THWN CAJAS NECESARIAS, INTERRUPTOR DE ACCIONAMIENTO , ALAMBRE DE COBRE Nº 12 AWG THWN Y DEMÁS ELEMENTOS NECESARIOS DESDE CADA LÁMPARA, PASANDO POR INTERRUPTORES HASTA EL TABLERO NORMAL DE PISO. UN CIRCUITO DE ILUMINACIÓN TENDRÁ MÁXIMO 1000 W INSTALADOS.)	1	1-5 días	\$ 82.565,08
3,02	INSTALACIONES ELÉCTRICAS. (SUMINISTRO DE MATERIALES EN GENERAL Y EJECUCIÓN DE SALIDA ELÉCTRICA PARA CIRCUITOS EXCLUSIVOS DE IMPRESORAS, FOTOCOPIADORA, RACK SE INCLUYE TUBERIA METALICA NECESARIA , REGATAS, CAJAS NECESARIAS, CABLE DE COBRE TRENZADO THWN Nº 12 AWG, COLORES SEGÚN NORMA, RESANES Y DEMÁS ELEMENTOS NECESARIOS DESDE EL TABLERO DE PISO DE LA RED NO REGULADA HASTA CADA PUNTO. EL SISTEMA ES MONOFÁSICO INCLUYE LÍNEA A TIERRA.. LAS TOMAS DEBEN SER ROTULADAS CON EL NÚMERO DEL CIRCUITO CORRESPONDIENTE).	1	1-5 días	\$ 89.452,92

Tabla N°4 Costos de Iluminación Y Tomas

8.6 EQUIPOS DE CÓMPUTO Y SERVIDOR

4,00	EQUIPOS DE COMPUTO Y SERVIDOR			
4,01	RACKS "Gabinete Alto 29 x Ancho 52 Fondo x 31 cm medidas externas,(5RU) "	7	-	\$ 1.050.000,00
4,02	COMPUTADORES DE GESTION DE DATOS " Procesadores Intel Xeon E5-2420 v2 ,Memoria de 8 GB, Dos discos duros de 1TB"	5	-	\$ 17.500.000,00

4,03	CONMUTADORES " Se recomienda switches de la serie C9000 de Dell Networking"	8	-	\$ 12.000.000,00
4,04	SEGURIDAD FIREWALL (Se recomienda Fortinet Fortigate 60d - Forticare Y Fortiguard Utm Firewall de cisco)	3	-	\$ 11.160.000,00
4,05	DISCO DURO EXTERNO 1 TB	5	-	\$ 875.000,00

Tabla N°5 Costos de Hardware

8.7 CAPACITACIÓN Y CIENTÍFICOS DE DATOS

5,00	CAPACITACIONES Y CIENTIFICOS DE DATOS			
5,01	CAPACITACION EN MANEJO Y GESTION DE REDES (DIRECTRICES OPERATIVAS DE LA EMPRESA Y RECURSOS HUMANOS)	2	3-5 días	\$ 2.100.000,00
5,02	CONTRATACION DE CIENTIFICO DE DATOS (DEBERA ESTAR CERTIFICADO COMO TAL, Y SU SALARIO NO EXCEDERA 2 VECES EL VALOR DE UN INGENIERO DE SOPORTE)	2	1,00	\$ 5.200.000,00

Tabla N°6 Costos de Capacitación

8.8 PLANOS Y CERTIFICACIONES

6,00	PLANOS Y CERTIFICACIONES			
6,01	SERVICIOS DE DIBUJO DE ARQUITECTURA (SUMINISTRO DE MANO DE OBRA Y MATERIALES PARA LA ELABORACIÓN DE LOS PLANOS ELECTRICOS (2) EN LOS CUALES SE INDIQUE CLARAMENTE LA DISTRIBUCIÓN FINAL DE LAS INSTALACIONES ELECTRICAS DE TOMAS Y FUERZA (1) Y DE ILUMINACIÓN (1). INCLUYE ELABORACIÓN DE DIAGRAMAS UNIFILARES DE LOS TABLEROS Y MARCACIÓN DE LOS MISMOS. SE DEBEN ENTREGAR EN COPIA DURA Y MEDIO MAGNÉTICO).		--	\$ 769.230,77
4,02	SERVICIOS DE ACREDITACIONES (HONORARIOS CAUSADOS POR LA OBTENCIÓN DEL CERTIFICADO DE CONFORMIDAD DE LA INSTALACIÓN ELÉCTRICA, EXPEDIDA POR UNO DE LOS ORGANISMOS ACREDITADOS ANTE LA SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO PARA TAL FIN, DE ACUERDO A LO DISPUESTO EN EL ARTÍCULO 47 DEL REGLAMENTO TÉCNICO DE INSTALACIONES ELÉCTRICAS (RETIE) Y A LA CIRCULAR 18 012 DE MARZO 2 DEL 2007 DEL MINISTERIO DE MINAS Y ENERGÍA.		--	\$ 1.117.647,06

4,02	SERVICIOS DE ACREDITACIONES (HONORARIOS CAUSADOS POR LA OBTENCIÓN DEL CERTIFICADO DE CONFORMIDAD DE LA INSTALACIÓN ELÉCTRICA Y DE REDES, EXPEDIDA POR UNO DE LOS CINCO ORGANISMOS ACREDITADOS ANTE LA ONAC PARA TAL FIN, DE ACUERDO A LO DISPUESTO EN EL REGLAMENTO PUBLICO (RETILAP), Y RESOLUCION 181331 DE AGOSTO DE 2009 DEL MINISTERIO DE MINAS Y ENERGÍA.		GLB.	\$ 1.117.647,06
------	---	--	------	-----------------

Tabla N°7 Costos de planos y certificaciones

* De lo cual se obtiene el total de costos que se presenta a continuación:

VALOR COSTOS IMPLEMENTACION DE BIG DATA PARA PYMES		
SUBTOTAL		\$ 58.117.779,48
Administración	10 %	\$ 5.811.777,95
Imprevistos	6%	\$ 3.487.066,77
TOTAL		\$ 67.416.624,20

Tabla N°8 Costo Total

Se habló de manera general de los costos que implica toda la estructura que compone la acometida eléctrica, los sistemas de seguridad, los soportes energéticos, entre otros, pero la capacitación en Big Data para que el personal maneje de manera adecuada la información no se contempló anteriormente y este es un parámetro muy influyente en el éxito de la innovación, además se habló de manera general de los costos estimados según la entidad datacenter capital cost, sin embargo una manera más específica de abordar el análisis implica contemplar soluciones ofrecidas por los prestadores de servicio que varían respecto a nivel de seguridad, capacidad de procesamiento, y costos. Cabe recordar que los costos anteriormente presentados fueron presentados pensando en un pequeño centro de datos para que la PYME de comunicaciones pudiera implementar soluciones como cloud computing o data analytics corrientes que son parte integral de una solución de Big Data.

El costo total obtenido de \$67.416.624,20, es un valor aproximado del valor a considerar para una inversión de una pequeña o mediana empresa que quiera implementar una solución de Big data, los precios de los equipos y materiales aquí presentados fueron proporcionados mediante cotizaciones con diferentes proveedores en la ciudad de Bogotá, que son citados en la sección de agradecimientos.

CAPITULO 9

CONCLUSIONES

Al desarrollar el presente trabajo, se quiere ampliar el campo de visión sobre los distintos conceptos que abarcaban la ideología de Big Data. Se han revisado los diversos enfoques de las implementaciones a nivel mundial que se han hecho sobre esta temática, en adición se han citado los proyectos en el país que están en desarrollo, los indicadores de crecimiento para el sector TIC y el campo de inclusión que tienen las pymes en este, además de proponer un modelo descriptivo para la implantación de iniciativas de Big Data y un esquema de costos en aras de presupuestar los factores económicos incidentes. A continuación, se exponen un conjunto de conclusiones específicas que constituye los resultados de esta monografía.

- Deberá entenderse por Big Data un compendio de tecnologías, una corriente evolutiva desde el punto de vista de las TICs, pues no es una tecnología como tal, sino un arreglo integral que abarca desde la captura de datos masiva, su almacenaje hasta su interpretación, este concepto deberá ser claro para la organización al momento de tomar la decisión de implementar soluciones Big Data.
- Se comprende que es viable implementar Big Data en Colombia para PYMES, pues su alcance se ajusta de manera perfecta a los indicadores de crecimiento que Con respecto al año anterior, el PIB cerró en el segundo trimestre de 2015 en 4,2% y la actividad de Correo y Telecomunicaciones en 4,8%, con una participación en el PIB de 3,2%, además de la amplia necesidad que se tiene por mantener las tendencias innovadoras en el mercado.
- Se realizó un estudio comparativo entre las diferentes ofertas en el mercado que manejan un gran volumen de datos, de esto se pudo concluir que para una PYME la mejor opción es usar las plataformas que ofrecen cloud computing y data analytics para el caso colombiano SAS o Hp Vertica pues ofrecen soluciones entre 800 y 1500 dólares que comparadas con una inversión de un valor aproximado a los 60 millones de pesos, son mucho más rentables.
- Deberá entenderse a los factores que inciden en el costo de una iniciativa de Big Data como directamente proporcionales a las necesidades que se quieran cubrir pues el modelo de costos que se propuso es orientado a suplir las necesidades de una PYME de comunicaciones, un sector diferente al de las TICs requerirá por supuesto suplir necesidades diferentes que inferirán directamente en el costo de la iniciativa.
- Se propuso un modelo descriptivo para la implementación de soluciones Big Data en PYMES consistente en 10 pasos a seguir, considerando cuales eran las etapas más

importantes a tener en cuenta dadas todas las implementaciones a nivel mundial y basándose en los casos de éxito de Big Data y contrastándolo con el modelo de Kevin C. Desouza de la universidad estatal de Arizona.

- Dado el continuo avance de la información y las cantidades de datos generadas constantemente crecientes en el mundo, el Big data debe continuar evolucionando para hacer frente a estos volúmenes de datos cada vez mayores e integrando sistemas que permitan diferenciar los datos del ruido, entre otros para optimizar el espacio de almacenamiento disponible y la velocidad de procesamiento de los datos.
- A pesar de que el volumen de datos que maneja la tecnología Big data es considerable, resulta no ser el más lucrativo para las compañías, puesto que esta inmensa cantidad de información la manejan pocas entidades a nivel global, razón por la cual la inclinación de las empresas para utilizar esta tendencia ha de ser en gran medida la variabilidad y la velocidad en el manejo de datos lo cual permite a las compañías seleccionar clientes potenciales, tendencias y publicidad rápidamente para multiplicar sus ganancias, no obstante no es la opción más adecuada en todos los casos.
- El análisis de costos que se realizó arrojó como resultado un valor cercano a los 67 millones de pesos este valor es un acercamiento al presupuesto que podría manejar una PYME si quisiera implementar soluciones de Big Data con infraestructura incorporada dentro de la organización.
- Una de las mayores implicaciones en cuanto al uso de la tecnología Big Data radica en los elevados costos que requiere una infraestructura para la correcta operación de los datos tal y como se evidenció en el análisis de costo para PYMES anteriormente, razón por la cual resulta conveniente aplicar una solución como la que ofrece el almacenamiento de datos en la nube a través de proveedores, con lo cual se reducen los costos y se obtiene una solución al inconveniente generado por la infraestructura.
- Luego de realizar el análisis de la información recolectada se llega a la conclusión de que es mejor contar con la asesoría de un experto en Big Data si se decide trabajar con la tecnología para evitar grandes fracasos o errores y permitiendo así la optimización del negocio.
- Dado que se trabajan volúmenes de datos exorbitantes (Orden de Petabytes o superior) y de distintas procedencias es uno de los puntos más notables del Big Data su análisis predictivo, el cual permite aprovechar toda esta información para generar sugerencias y suposiciones de las preferencias de los clientes y es este análisis predictivo lo que lo hace tan rentable para las PYMES puesto que evita costos de análisis de mercadeo al hacerse una idea acertada de las tendencias del comportamiento humano basado en los lineamientos con los que se alimenta el sistema, desde sus inicios en la organización otorgando así productos y servicios adaptados a las necesidades de los usuarios.

CAPITULO 10

BIBLIOGRAFIA

- [1] EMC (2010). Artículo especial de "The Economist" <http://www.emc.com/collateral/analystreports/ar-the-economist-data-data-everywhere.pdf>
- [2]. Boletín trimestral de las TIC - Cifras primer trimestre de 2015. www.mintic.gov.co/portal/604/w3-article-6163.html
- [3]. Lal K, Mahanti N C. A Novel Data Mining Algorithm for Semantic Web Based Data Cloud[J]. International Journal of Computer Science and Security (IJCSS), 2010, 4(2): 160.
- [4]. <https://www.renata.edu.co/index.php/convocatorias/7866-encuentro-mundial-de-big-data-colombia-2015>
- [5].Brad Helund, Understanding Hadoop Cluster and the network, <http://bradhedlund.com/2011/09/10/understanding-hadoop-clustersand-the-network/>.
- [6]. Tom White, Hadoop: The Definitive Guide (second edition[M], O'Reilly Press, 2010:15-38
- [7]. Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters <http://labs.google.com/papers/mapreduce.html>
- [8].Hadoop Poses a Big Data Security Risk: 10 Reasons Why, <http://www.eweek.com/security/slideshows/hadoop-poses-a-big-datasecurity-risk-10-reasons-why>
- [9].Shastri L Nimmagadda and Heinz Dreher (2011) Data warehousing and mining technologies for adaptability in turbulent resources business environments, *Int. J. Business Intelligence and Data Mining*, Vol. 6, No. 2, 2011, p 113-153
- [10]. www.cisco.com/web/ES/pdf/Cisco_ASR_2012_v2_020813.pdf
- [11]. <http://datos.bancomundial.org/indicador/IT.NET.USER.P2/countries/1W?display=default>
- [12]. Referenced by Nathan Eagle in video interview for UN Global Pulse, July 2011. Though, the term seems to have been originally coined by Joe Hellerstein, a computer scientist at the University of California,Berkeley <<http://www.economist.com/node/15557443>>
- [13]. Onella, Jukka- Pekka. "Social Networks and Collective Human Behavior." UN Global Pulse. 10 Nov.2011. <<http://www.unglobalpulse.org/node/14539>>Lohr, Steve. "The Age of Big Data." New York Times. 11 Feb, 2012.<http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=2&pagewanted=all>
- [14]. Kirkpatrick, Robert. "Digital Smoke Signals." UN Global Pulse. 21 Apr. 2011. <<http://www.unglobalpulse.org/blog/digital-smoke-signals>>
- [15]. "The Data Deluge." The Economist. 25 Feb 2010. <<http://www.economist.com/node/15579717>> andAmmirati, Sean. "Infographic: Data Deluge – 8 Zettabytes of Data by 2015." Read Write Enterprise.
- [16.] <<http://www.readwriteweb.com/enterprise/2011/11/infographic-data-deluge---8-ze.php>>
- [17]. King, Gary. "Ensuring the Data-Rich Future of Social Science." Science Mag 331 (2011) 719-721. 11 Feb, 2011 Web. <http://gking.harvard.edu/sites/scholar.iq.harvard.edu/files/gking/files/datarich_0.pdf>
- [18]. Helbing, Dirk , and Stefano Balietti. "From Social Data Mining to Forecasting Socio-Economic Crises." Arxiv (2011) 1-66. 26 Jul 2011 <http://arxiv.org/pdf/1012.0178v5.pdf>.
- [19]. Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angela H. Byers. "Big data: The next frontier for innovation, competition, and productivity." McKinseyGlobal Institute (2011): 1-137. May 2011. <http://www.mckinsey.com/mgi/publications/big_data/pdfs/MGI_big_data_full_report.pdf>
- [20]. " World Population Prospects, the 2010 Revision." United Nations Development Programme. <http://esa.un.org/unpd/wpp/unpp/panel_population.htm>
- [21]. "Data Exhaust." <<http://www.wordspy.com/words/dataexhaust.asp>>

- [22]. Cornu, Celine. "Mobil Banking' Moving Through Developing Countries." Jakarta Globe. 21 Feb, 2010. <<http://www.thejakartaglobe.com/business/mobile-banking-moving-through-developingcountries/359920>>
- [23]. "Global Internet Usage by 2015 [Infographic]." Alltop. <<http://holykaw.alltop.com/global-internetusage-by-2015-infographic?tu3=1>>
- [24]. Rao, Dr. Madanmohan. "Mobile Africa Report: Regional Hubs of Excellence and Innovation." MobileMonday (2011): 1-68. Mar. 2011. <http://www.mobilemonday.net/reports/MobileAfrica_2011.pdf>
- [25]. "Big Data, Big Impact: New Possibilities for International Development." World Economic Forum (2012): 1-9. Vital Wave Consulting. Jan. 2012 <<http://www.weforum.org/reports/big-data-big-impact-newpossibilities-international-development>>.
- [26]. Artículos del 26 al 36 fueron proporcionados por Andrea Rodríguez y Rafael Niño , enlaces para el proyecto en el MINTIC y el DNP respectivamente, si se desea algún indicador en especial podrá ser solicitado al correo bigdataparapymecol@gmail.com.
- [27]. Eke, P.I.. "Using Social Media for Research and Public Health Surveillance." (Abstract). Journal of Dental Research 90.9 (2011). <<http://jdr.sagepub.com/content/early/2011/07/15/0022034511415273>>
- [28]. Signorini, Alessio, Alberto M. Segre, and Phillip M. Polgren. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic." PLoS One 6.5
- [29]. Moreno, Megan A., Dimitiri A. Christakis, Katie G. Egan, Libby N Brockman and Tara Becker.
- [30]. Realizing the Promise of Big Data Implementing Big Data Projects. Kevin C. Desouza Arizona State University, http://observgo.quebec.ca/observgo/fichiers/26986_Realizing%20the%20Promise%20of%20Big%20Data.pdf
- [31]. Students." Archives of Pediatrics and Adolescent Medicine (2011). <<http://archpedi.amaassn.org/cgi/content/abstract/archpediatrics.2011.180v1?maxtoshow=&hits=10&RESULTFORMAT=&fulltext=facebook&searchid=1&FIRSTINDEX=0&resourcetype=HWCIT>>.
- [32]. Walsh, Bryan. "Outbreak.com: Using the Web to Track Deadly Diseases in Real Time." Time Science. 16 Aug. 2011. <<http://www.time.com/time/health/article/0,8599,2088868,00.html>>.
- [33]. Helbing and Ballester. "From Social Data Mining to Forecasting Socio-Economic Crisis." The European Physical Journal-Special Topics. (Volume 195, Number 1, 3-68, pg. 24) 26 July 2011. <<http://arxiv.org/abs/1012.0178>>
- [34]. "Data Philanthropy: A New Frontier?" Rally the Cause. 2 Mar. 2011.
- [35]. <<http://rallythecause.com/2011/03/02/data-philanthropy-a-new-frontier/>>
- [36]. King, Gary N. and Eleanor Powell. How Not to Lie Without Statistics. Working Paper. Harvard University, 22 Aug. 2008. <<http://gking.harvard.edu/gking/files/nolie.pdf>>
- [37]. Pang, Bo and Lillian Lee. "Opinion Mining and Sentiment Analysis." Foundations and Trends in Information Retrieval 2.1-2 (2008): 1-135. NOW. <<http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf>>

AGRADECIMIENTOS

Nos gustaría que estas líneas sirvieran para expresar el más profundo y sincero agradecimiento a todas aquellas personas que con su ayuda han colaborado en la realización del presente trabajo, en especial al Ing. José David Cely Callejas, director de esta investigación, por la orientación, el seguimiento y la supervisión continúa de la misma, pero sobre todo por la motivación y el apoyo recibido a lo largo de estos meses.. También nos gustaría agradecer la ayuda recibida del profesor Gustavo Adolfo Higuera castro. Quisiéramos hacer extensiva nuestra gratitud a las compañías IT&T Ing. De Colombia S.A , SAS Latinoamérica, Serínsy “ Una nueva Visión “ y CONTRONET Carrera 10 , quienes fueron de vital importancia en el proceso de adquisición de cotizaciones información y demás aportes realizados a este proyecto, También queremos dar las gracias a Ivan Javier Sepulveda arquitecto quien proporción información detallada en el proceso de planos y certificaciones . Un agradecimiento muy especial merece la comprensión, paciencia y el ánimo recibidos de nuestros familiares y amigos.