PLAN DE IMPLEMENTACIÓN DE TECNOLOGÍAS BIG DATA PARA LA OPTIMIZACIÓN DE ESTRATEGIAS COMERCIALES Y DE SEGMENTACIÓN

Francisco Carrillo Álvarez

Universidad Autónoma de Bucaramanga
Facultad de Ingeniería de Sistemas
Maestría en Gestión, Aplicación y Desarrollo de Software
Bucaramanga
2016

PLAN DE IMPLEMENTACIÓN DE TECNOLOGÍAS BIG DATA PARA LA OPTIMIZACIÓN DE ESTRATEGIAS COMERCIALES Y DE SEGMENTACIÓN

FRANCISCO CARRILLO ÁLVAREZ

TESIS PARA OPTAR AL TÍTULO:
Magister en Gestión Aplicación y Desarrollo de Software

DIRIGIDO POR:
Director: MSc DIEGO FABIÁN PAJARITO GRAJALES
Co-Director: MSc MARITZA LILIANA CALDERON BENAVIDES

Universidad Autónoma de Bucaramanga
Facultad de Ingeniería de Sistemas
Maestría en Gestión, Aplicación y Desarrollo de Software
Bucaramanga
2016

Bucaramanga 08 de julio de 2016

Señores

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA

Ciudad

Asunto: Documento de Aprobación de Tesis

Yo, **Diego Fabián Pajarito Grajales**, identificado con cedula de ciudadanía **No 80.111.111 de Bogotá**, en calidad de director de la tesis de grado titulada "Plan de Implementación de Tecnologías Big Data para la Optimización de Estrategias Comerciales y de Segmentación", elaborada por el ingeniero **Francisco Carrillo Álvarez**, candidato al grado del programa Maestría en Gestión, Aplicación y Desarrollo de Software de la facultad de ingeniería de sistemas; doy constancia del cumplimiento de los objetivos propuestos, calidad, coherencia, buena redacción y uso de la terminología apropiada en la tesis presentada, y emito el concepto de ***TESIS APROBADA.***

La presentación del trabajo se realizó el día 20 de junio de 2016, asistiendo como calificadores los ingenieros Miguel Antonio Cadena Carter y Claudia Isabel Cáceres Becerra.

Atentamente,

Diego Fabián Pajarito Grajales

Director

Nota de Aceptación:

Firma Presidente del Jurado.

Firma del Jurado.

Firma del Jurado.

Bucaramanga, Junio 20 de 2016

A DIOS por las fuerzas y la constancia para terminar esta tesis.
A mi hija Salomé por esas horas de atención y juego que sacrifique.
A mi esposa por la comprensión y apoyo.
A mi tía María, Madre y Hermanos por el apoyo y preocupación.
A mis amigos y compañeros de trabajo por su motivación.

AGRADECIMIENTOS

En todo trabajo de investigación participan personas e instituciones, que gracias a su participación llega a feliz término.

Esta, tesis de maestría, no es una excepción. Por consiguiente deseo expresar mi más sincero agradecimiento a quienes directa e indirectamente han participado en su desarrollo.

En primer lugar agradezco a mi director de tesis el MSc. Diego Fabián Pajarito Grajales, por su compromiso y disposición en todo momento. Además de su conocimiento y guía invaluables para dar termino a esta tesis.

De manera especial expreso mi más sincero agradecimiento a la Co Directora de tesis, MSc. Maritza Liliana Calderón Benavides. Por sus consejos recomendaciones y aportes en el desarrollo de este proyecto.

Finalmente agradecimientos especiales a Vanguardia Liberal por el apoyo económico brindado para llevar acabo los estudios de la maestría en Gestión, Aplicación y Desarrollo de Software. De la misma forma agradecimientos especiales al Ing. René Di marcó Sub Gerente de Tecnología y a la Doctora Luz Helena Rodríguez Sub Gerente de Gestión Humana, por los aportes, espacios, permisos y gestión valiosos para dar término a los estudios adelantados.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ANEXOS

RESUMEN

La analítica de datos masivos ha disparado las iniciativas hacia la implementación de aplicaciones de *Big Data* en muchas organizaciones de todos los sectores económicos del país. La planeación y despliegue de estrategias comerciales exitosas, procesos adecuados de segmentación de clientes, implementación de políticas de fidelización, gestión del conocimiento y la unificación de la información son algunos de los aspectos más relevantes para empresas que buscan mejorar su desempeño comercial a través de la explotación de su activo más importante –la información-. Los medios impresos, por su naturaleza, son constantemente llamados a implementar herramientas que le permitan mejorar no solo el conocimiento del cliente sino sus procesos de gestión de información. En el presente estudio se hace una revisión de las necesidades de adopción de nuevas fuentes de información en procesos de venta, transformación y unificación de las diferentes fuentes de datos existentes en la empresa, se plantea una arquitectura tecnológica que soporte dichos cambios basada en la implementación de tecnologías *Big Data*, modelos de procesos de unificación de la información y de ventas, y a su vez se propone un plan de implementación que brinde a Vanguardia Liberal las fortalezas necesarias para afrontar procesos de analítica de datos masivos en el sector de los medios impresos.

Palabras clave: *Big Data*, Estrategias comerciales, Periódicos Impresos, Arquitectura, plan, Segmentación

GLOSARIO

ACID: (*Atomicity, Consistency, Insolación, Durability*) conjunto de características o propiedades para las bases de datos. Atomicidad, consistencia, aislamiento y durabilidad.

API: Del inglés, *Application Programming* Interface. Conjunto de funciones y métodos que se ponen a disposición del usuario u otro software, como una capa de abstracción, para facilitar la interacción con un determinado software o tecnología.

B2B: *Business to Business* (negocio a negocio). Tipo de *e-commerce* en el que una empresa vende sus servicios o productos a otra empresa.

B2C: *Business to Consumer* (negocio a cliente). Tipo de *e-commerce* más tradicional, en el que una empresa vende sus servicios o productos a clientes individuales.

CMR: *Customer Relationship Management* (Manejo de las Relaciones con el Cliente).

***Cloud Computing*:** Entorno en el que es posible almacenar diferentes tipos de contenido o aplicaciones, sin tener que disponer de una infraestructura propia que lo mantenga. Es la posibilidad de utilizar servicios en la Red sin disponer de la estructura necesaria que hace falta para mantener y ofrecer este tipo de servicio.

Clúster: Conjunto de computadores interconectados entre sí que actúan como uno solo.

***Data Lake*:** Sistema de almacenamiento, normalmente distribuido en un conjunto de máquinas, que alberga una gran cantidad de datos en su formato nativo. Acostumbra a ser el sistema de almacenamiento en los sistemas *Big Data*.

***Data Warehouse*:** Base de datos corporativa que integra información de una o más fuentes distintas, para ser procesada con el fin de realizar análisis que ayuden a la toma de decisiones en la entidad en la que se utiliza.

DashBoards: Es una representación gráfica de los principales indicadores (KPI) que intervienen en la consecución de los objetivos de negocio, y que está orientada a la toma de decisiones para optimizar la estrategia de la empresa.

ERP: *Enterprise Resource Planning* (Planeamiento de los Recursos de la Empresa).

ETL: Del ingles, *Extract, Transform and Load*. Proceso de extracción de información de diferentes fuentes, para su posterior transformación para ser almacenada en el formato idóneo en el sistema de almacenamiento de información deseado.

***Framework*:** Es una estructura conceptual la cual sirve de soporte para facilitar el desarrollo de software.

GPL: (GNU *General Public License*) Licencia Pública General de GNU.

HDFS: (*Hadoop Distributed File System*) sistemas de archivos distribuido Hadoop.

JSON: Del inglés, *JavaScript Object Notation*. Formato ligero de intercambio de da- tos basado en el lenguaje de programación JavaScript, fácil de comprender y escribir para humanos y simple de generar e interpretar por las máquinas.

***Mapping*:** Dentición del lugar concreto donde se guarda cierta información en un sistema de almacenamiento.

NOSQL: (No *Structured Query Language*) bases de datos no relacionales.

***Open Source*:** Modelo de desarrollo de software que fomenta el acceso universal y desarrollo colaborativo de un código fuente, a través de licencias que permiten el uso, mordicación y redistribución de este código.

P2P: *Point to Point*. Tipo de *e-commerce* que se da cuando dos usuarios intercambian bienes, servicios o algún valor de forma directa, sin ningún intermediario.

RDBMS: (*Relational Database Management System*) sistema de Gestión de Base de Datos Relacional.

RDF: Del ingles, *Resource Description Framework*. Modelo de datos estándar para la representación de metadatos en forma de grafo. Extiende el sistema de enlaces de la Web utilizando URIs para definir las relaciones entre los dos extremos de un enlace (comúnmente llamadas tripletas).

SPARQL: Del ingles, SPARQL *Protocol And RDF Query Language*. Lenguaje de consulta semántico para bases de datos, capaz de consultar y manipular datos guardados en formato RDF.

SQL *Server*: Es un motor de base de datos tradicional SQL de modelo relacional de tecnología Microsoft.

***Streaming*:** Es una transmisión continua digital de multimedia a través de un red de computadoras.

TIC: Tecnologías de la Información y la Comunicación.

Vista 360: Se trata de una estrategia integral de administración para atender y responder a las necesidades de los clientes. Permite manejar, consolidar y analizar de manera sencilla gran cantidad de información crucial para la empresa

XML: (*Extensible Markup Language*): lenguaje de marcas extensibles, utilizado por algunas bases de datos

INTRODUCCIÓN

Actualmente las empresas de todos los sectores están preocupándose por como conocer y disponer de la información que tienen de sus clientes dentro de su misma empresa y en su propia infraestructura. El tema en este aspecto es como reunir toda esa información dispersa en múltiples sistemas de información y fuentes de datos estructuradas, semi estructuradas o no estructuradas, que puede entorpecer la gestión de la fuerza de ventas al no contar con la información adecuada para mantener las relaciones con otras áreas del negocio. Otros aspectos relacionados con el tratamiento de la información se ven afectados por la falta de integración como son: Procesos de segmentación y estrategias comerciales que dejan de ser óptimas al ser generadas o analizadas con información parcial.

Vanguardia Liberal no es ajena a los problemas antes mencionados. Por esta razón el presente trabajo de investigación busca diseñar un plan de implementación de tecnologías de *Big Data*, que permita dar solución a los inconvenientes de integración de múltiples fuentes de datos, eliminando el problema de las islas de información, por consiguiente mejorar los procesos de segmentación y estrategias comerciales de la empresa, disponer de información integrada y completa, para finalmente facilitar procesos de vista 360 de los clientes. Punto de partida de las empresas que basan sus decisiones en el análisis de datos. Esta tesis está planteada en el desarrollo de los siguientes capítulos así: Capitulo 1 Marco metodológico el cual describe las principales consideraciones que dan origen al presente trabajo, capitulo 2 Marco teórico que proporciona la fundamentación teórica que soporta el desarrollo de la investigación, capitulo 3 Diagnostico del estado actual en el que se explica la arquitectura actual del proceso de negocio y sus diferentes procesos de venta, capitulo 4 Identificación de requerimientos, en donde se referencia la forma en que se identifican los requerimientos de Gerentes, Sub Gerentes y Coordinadores de área de las unidades de negocio de publicidad, circulación, administración y tecnología. El capítulo 5 Tecnologías Big Data que menciona la selección de herramientas *Big Data*, el Capítulo 6 describe las arquitecturas y modelos de análisis propuestos, el Capítulo 7 presenta el plan de implementación de tecnologías *Big Data* y finalmente en el Capítulo 8 son presentadas las conclusiones de la presente tesis.

1. MARCO METODOLÓGICO

En este capítulo se plantea la problemática que afronta el periódico, se argumentan las razones que permiten identificar la implementación de tecnologías *Big Data* como una solución al problema, se plantean los objetivos de la investigación, la metodología a seguir y las etapas de la presente investigación.

1.1 Planteamiento del Problema

El periódico Vanguardia Liberal al igual que muchas empresas comerciales padece en la actualidad del síndrome denominado “Islas de información”(Coronel, 2011), que entorpece en cierta medida el conocimiento y la generación de una vista 360.

El manejo de múltiples fuentes de información, tanto estructurada como no estructurada, representa un inconveniente en aquellas organizaciones que no cuentan con tecnologías apropiadas que permitan la vinculación de esa información a los procesos de evaluación y toma de decisiones.

Los procesos de análisis de información con los que cuenta Vanguardia Liberal, representan costos y demoras en la toma de decisiones, la segmentación actual de los clientes no es adecuada para lograr los objetivos de las estrategias comerciales.

El procesamiento aislado de la información relacionada con el negocio desde la perspectiva individual de cada canal de venta de la organización, no permite la adecuada generación de datos para procesos analíticos que permitan conocer la dinámica empresarial. Actualmente la compañía no cuenta con herramientas para el adecuado procesamiento de datos del negocio que faciliten el análisis holístico de los datos generados desde las diferentes unidades de negocio de la compañía.

1.2 Justificación

En la actualidad hablar de *Big Data*, significa: Tomar decisiones más acertadas, proyectar estrategias comerciales más efectivas, mejorar la experiencia de cliente y aumentar la fidelización entre otros aspectos relacionados directamente con procesos comerciales. Todo esto gracias a que hoy se cuenta con mayores fuentes de información que constantemente crecen, las cuales son consideradas difíciles, complejas, lentas o costosas a la hora de integrarlas a procesos de análisis dentro de las organizaciones.

Los medios de comunicación y en particular los periódicos impresos, no son ajenos a las necesidades antes planteadas frente al mejoramiento de la relación con los clientes. Tener una vista de 360 grados que permita a la empresa conocer cuál es el comportamiento de los clientes por los diferentes canales de venta de la organización es hacia donde se están encaminando todos los esfuerzos.

Actualmente Vanguardia Liberal se encuentra llevando a cabo un plan que permitirá solucionar este inconveniente en pro de mejorar la relación con los clientes. El desarrollo de este estudio aporta un valioso conocimiento que establecerá políticas a seguir en un mediano plazo reforzando la iniciativa de la organización.

Este proyecto busca establecer un plan de implementación de tecnologías *Big Data* que permitirán a la empresa mejorar la toma de decisiones de tipo estratégico al contar con la evaluación de mayor cantidad de datos, además de contemplar dentro de sus análisis información que antes no se evaluaba. Con este proyecto se busca que la segmentación de sus clientes redunde en campañas de mercadeo mucho más efectivas y que produzcan una mejor “experiencia” para él.

Se ofrecerá un plan de implementación de tecnologías *Big Data* correctamente asociado a los procesos de negocio, haciendo uso de una adecuada selección de herramientas, para permitir a la empresa procesar los datos derivados del negocio y posteriormente generar información que soporte nuevos niveles de análisis para mejorar el conocimiento empresarial, lo cual a su vez generará grandes beneficios y mayor productividad dentro de los grupos de trabajo.

1.3 Pregunta De Investigación

¿Qué tecnologías de *Big Data*, son las más apropiadas para diseñar el plan de implementación que busca mejorar las estrategias comerciales y de segmentación al interior del periódico Vanguardia Liberal?

1.4 Objetivos De La Investigación

1.4.1 Objetivo General

Definir un plan de implementación de tecnologías "*Big Data*" para VANGUARDIA LIBERAL a partir del direccionamiento estratégico de la organización, con el fin de optimizar las estrategias comerciales y de segmentación.

1.4.2 Objetivos Específicos

- Identificar los activos de información candidatos a ser estructurados y analizados con tecnologías de *Big Data* que permitan optimizar las estrategias comerciales y de segmentación.

- Plantear modelos de análisis con técnicas de *Big Data* que permitan una mejor formulación de estrategias comerciales y de segmentación en Vanguardia Liberal.

- Evaluar las tecnologías actualmente disponibles en ámbitos de *Big Data* para soportar la formulación de estrategias comerciales y de segmentación.

- Evaluar la viabilidad técnica del plan de implementación de tecnologías *Big Data* acorde con el direccionamiento estratégico de Vanguardia Liberal.

1.5 Resultados Esperados

- Inventario de activos de información cuyo análisis en ámbitos *Big Data* permite optimizar las estrategias comerciales y de segmentación.

- Modelos de análisis propuestos para soportar estrategias comerciales y de segmentación al interior de Vanguardia Liberal.

- Selección y evaluación de las tecnologías *Big Data* adecuadas para la formulación de estrategias comerciales y segmentación en Vanguardia Liberal.

- Plan de implementación de tecnologías *Big Data* para la formulación de estrategias comerciales y de segmentación en Vanguardia Liberal.

- Recomendaciones técnicas y lineamientos de Política de gestión de información para soportar estrategias comerciales y procesos de segmentación acorde con el direccionamiento estratégico de Vanguardia Liberal

1.6 Metodología

La metodología utilizada en el desarrollo de este trabajo estuvo basada en el modelo de prototipos, el cual se centra en una representación de aquellos aspectos que sean visibles para el cliente o usuario final, construyendo prototipos que permiten la retroalimentación de los interesados que facilitan el refinamiento

de los requisitos iniciales planteados y a su vez garantiza un entendimiento pleno del caso de estudio.(Lawrence, 2002)

El uso de prototipos se ha utilizado en Vanguardia Liberal en los últimos 10 años, dada su facilidad de uso y las ventajas que implica la retroalimentación temprana de los interesados del negocio, al poder validar los resultados esperados o las dificultades que se presentan durante el desarrollo de cada uno de los prototipos, que bien pueden obedecer a requerimientos cortos de alcance en su definición inicial.

Un caso reciente fue la implementación de una nueva política de comisiones planteada para la fuerza de ventas que consistía en generar un ingreso más estable y permanente, favorable tanto para la empresa como para los asesores pertenecientes a la unidad de negocio de Publicidad.

Las ventajas de la aplicación de modelos de prototipos en este caso permitió definir, tres iteraciones donde se evidenciaron fallas de alcance en el planteamiento inicial de los requerimientos que permitieron hacer los ajustes necesarios y a su vez se mejoraron en cada uno de los modelos la forma de visualizar los resultados.

La principal desventaja de la utilización del modelo de prototipos, experimentada en este desarrollo puntual fue el cambio de los requerimientos iniciales respecto al alcance del desarrollo planteado, en dos de las tres iteraciones realizadas durante los dos meses de desarrollo.

Basados en las experiencias que la organización ha tenido en el modelado de prototipos, se determina que esta, es la metodología más adecuada para adelantar la presente investigación.

1.7 Etapas de la Investigación

En la tabla 1, se describen las etapas que se abordaron en el desarrollo de la investigación.

1.8 Actividades

En la tablas 2, 3, 4 y 5 se detallan los objetivos del proyecto y la forma como se pretendieron alcanzar, mencionando las actividades, las fuentes de las cuales se obtendrán la información, las técnicas que se emplearán y los respectivos entregables de cada uno de estos.

Tabla 1: Etapas de la Investigación

Etapa	Descripción	Objetivo Mapeado
Identificación de Activos de Información	A partir de la revisión del concepto de activo de información, se identifican y clasifican los diferentes activos de Información al interior de la empresa	Identificar los activos de información candidatos a ser estructurados y analizados con tecnologías de *Big Data* que permitan optimizar las estrategias comerciales y de segmentación.
Análisis de Activos de Información	En base al levantamiento realizado de los diferentes activos de información, se procede con la asignación del peso representativo a cada uno de ellos, el cual identificará los nichos más importantes de información con los que cuenta la empresa.	Plantear modelos de análisis con técnicas de *Big Data* que permitan una mejor formulación de estrategias comerciales y de segmentación en Vanguardia Liberal.
Investigación sobre el Ecosistema de Tecnologías de *Big Data*	A partir de la revisión bibliográfica de las diferentes Tecnologías de *Big Data*, más representativas, proponemos las más adecuadas para el tipo de activos de información que fueron detectados	Evaluar las tecnologías actualmente disponibles en ámbitos de *Big Data* para soportar la formulación de estrategias comerciales y de segmentación.
Análisis de las tecnologías *Big Data* propuestas	Basados en los requerimientos técnicos particulares de cada una de las herramientas propuestas, Soporte, tamaño de la comunidad que se identifica con las herramientas, se harán recomendaciones	Evaluar la viabilidad técnica del plan de implementación de tecnologías *Big Data* acorde con el direccionamiento estratégico de Vanguardia Liberal.

Fuente: Autor

Tabla 2: Primer Objetivo Específico del Proyecto

Objetivo	Identificar los activos de información candidatos a ser estructurados y analizados con tecnologías de *Big Data* que permitan optimizar las estrategias comerciales y de segmentación.		
Actividades	**Técnicas**	**Fuentes**	**Entregables**
• Levantamiento del inventario actual de los activos de información existentes en la empresa.	• Encuesta / Levantamiento de requerimientos*	• Información suministrada por Sub Gerentes y Coordinadores de las diferentes unidades de la empresa.	• Inventario de activos de información que potencialmente serían analizados con tecnologías *Big Data*

Fuente: Autor

** Pese a no ser un proceso formal dentro de un proceso de ingeniería de software se espera que los stakeholders aporten la mayoría de las necesidades del negocio para identificar cuáles y de qué forma pueden ser soportadas con tecnologías Big data.*

Tabla 3: Segundo Objetivo Específico del Proyecto

Objetivo	Plantear modelos de análisis con técnicas de *Big Data* que permitan una mejor formulación de estrategias comerciales y de segmentación en Vanguardia Liberal.		
Actividades	**Técnicas**	**Fuentes**	**Entregables**
• Evaluación y clasificación del Inventario de Activos de Información.	• Encuesta / Levantamiento de requerimientos*	• Información suministrada por Sub Gerentes y Coordinadores de las diferentes unidades de la empresa.	• Modelos de análisis y aprovechamiento propuestos para soportar estrategias comerciales al interior de Vanguardia Liberal.

Fuente: Autor

** Pese a no ser un proceso formal dentro de un proceso de ingeniería de software se espera que los stakeholders aporten la mayoría de las necesidades del negocio para identificar cuáles y de qué forma pueden ser soportadas con tecnologías Big data.*

Tabla 4: Tercer Objetivo Específico del Proyecto

Objetivo	Evaluar las tecnologías actualmente disponibles en ámbitos de *Big Data* para soportar la formulación de estrategias comerciales y de segmentación.		
Actividades	**Técnicas**	**Fuentes**	**Entregables**
• Investigar sobre las diferentes herramientas existentes que conforman el ecosistema de tecnologías de *Big Data*	• Revisión Documental	• Documentos publicados en revistas científicas • Artículos publicados en revistas • Libros relacionados con el tema	• Identificación de las tecnologías existentes aplicables para proyectos *Big Data* en Vanguardia Liberal
• Ponderación de beneficios y desventajas que ofrecen las tecnologías propuestas en el marco de la organización	• Ponderación • Consulta a expertos, o *stakeholders*	• Documentos publicados en revistas científicas • Artículos publicados en revistas • Libros relacionados con el tema	• Evaluación de las tecnologías adecuadas para la implementación de proyectos *Big Data* en Vanguardia Liberal

Fuente: Autor

Tabla 5: Cuarto Objetivo Específico del Proyecto

Objetivo	Evaluar la viabilidad técnica del plan de implementación de tecnologías *Big Data* acorde con el direccionamiento estratégico de Vanguardia Liberal.		
Actividades	**Técnicas**	**Fuentes**	**Entregables**
• Evaluar los requerimientos de las tecnologías vs los elementos disponibles en la empresa.	• Revisión Documental	• Documentos publicados en revistas científicas • Artículos publicados en revistas • Libros relacionados con el tema	• Propuesta inicial del Plan de implementación, análisis y aprovechamiento de grandes volúmenes de datos dentro de Vanguardia Liberal.
• Dimensionar tiempos de implementación • Formular plan de implementación, de las tecnologías *Big Data.*	• Revisión Documental • Validación con la dirección	• Documentos publicados en revistas científicas • Artículos publicados en revistas • Libros relacionados con el tema	• Recomendaciones y propuesta de Políticas para gestionar la información de Clientes de los diferentes canales comerciales de la empresa, que faciliten la comprensión de necesidades e intereses.

Fuente: Autor

2. MARCO TEÓRICO

Según los estudios publicados por (Selvage, 2013) y (Buytendijk & Laney, 2014) (Morales & Carolina, 2015).“Las empresas no manejan un modelo de integración de la información y gestión del conocimiento; los datos se encuentran dispersos en documentos, bases de datos, redes sociales, etc., esto ha generado la existencia de islas de datos independientes que en algunas ocasiones limitan la visión de una empresa para la toma de decisiones y a esta complicación se suma la explosión del volumen de información para manejarla adecuadamente tanto en datos estructurados como no estructurados” Estos estudios son corroborados por la situación actual a la que se enfrentan las organizaciones.

El sector de los medios de comunicación en concreto el de los medios impresos, no es ajeno al desafío que representa en la actualidad la manipulación, administración, almacenamiento, búsqueda y análisis de grandes volúmenes de datos. Una de las tecnologías que ha entrado en el panorama mundial en los últimos años con gran fuerza y que permite a las organizaciones afrontar este gran reto es: *Big Data*.

En los siguientes apartados se resuelven interrogantes como: ¿Qué es *Big Data*?, ¿Qué es un Cliente?, ¿Qué es Segmentación?, ¿Qué es una Estrategia empresarial? y se define el concepto de Activos de Información, importantes todos ellos para conocer y entender el significado, la relevancia y relación que toman dentro del desarrollo del proyecto. Además se dimensiona de qué forma la implementación de tecnologías *Big Data* puede mejorar la relación con clientes mediante la optimización de estrategias comerciales y de segmentación, utilizando sus propios activos de información.

2.1 *Big Data*

Muchas de las innumerables definiciones sobre *Big data* centran su conclusión sobre el paradigma de las 3vs, volumen, velocidad, variedad y en los últimos años ha surgido una cuarta “V”, que hace referencia a la veracidad de los datos(López García, 2013), enfocado a la incapacidad de procesar esta ingente cantidad de información en los sistemas tradicionales.

Una de las aproximaciones más completas de *Big Data* con relación a un entorno empresarial es la suministrada según Gartner (Goyzueta Rivera, 2015). “Son

activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones." Este trabajo se identifica plenamente con la definición propuesta por Gartner, que define el término de *Big Data* para los grandes conjuntos de datos y al conocimiento que puede arrojar el análisis de los mismos.

2.1.1 Las 3 "V"

A continuación se presenta el significado de cada una de las 3 "V", que consideran muchos autores como el paradigma del *Big Data*:

Volumen: Se relaciona con la cantidad de información recolectada, que incluye Información de fuentes tradicionales y no tradicionales. Las empresas trabajan ahora con *PetaBytes* (10^{15} *bytes*) y *ExaBytes* (10^{18} *bytes*)(Morales & Carolina, 2015). En Vanguardia Liberal, un cliente puede llegar a ser suscriptor del periódico impreso, anunciante de avisos publicitarios, comprador frecuente de artículos optativos (Carros de Colección, Utensilios de Cocina, etc.) y cliente de avisos clasificados entre otros roles. Esta Información relacionada genera valor al interior de la empresa.

Velocidad: Se refiere a la velocidad con la que la información es generada y fluye hacia la empresa(Morales & Carolina, 2015). Un análisis realizado por la empresa Domo, en el año 2012 sobre la cantidad de información que los internautas dan de uso a la red cada minuto indica que: "Cada minuto que pasa, los 2.700 millones de personas con acceso a Internet que se calcula que hay actualmente en el mundo, envían más de 200 millones de correos electrónicos, realizan 2 millones de consultas a *Google*, suben 48 horas de vídeo a *YouTube*, escriben más de 100.000 mensajes en *Twitter*, publican casi 30.000 nuevos artículos en sitios como *Tumblr* o *WordPress*, suben más de 6.000 fotografías a *Instagram* y *Flickr*, se descargan 47000 aplicaciones del sistema operativo IOS…"(López García, 2013)

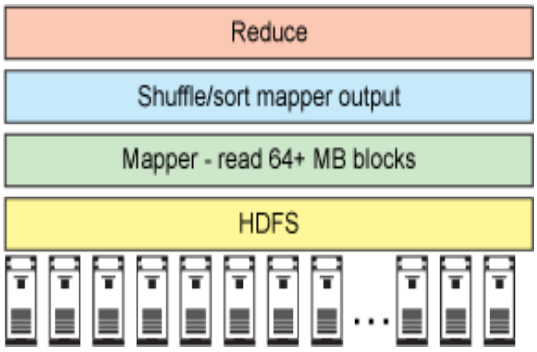
Variedad: Se refiere al tipo de información disponible para la empresa y para sus equipos de marketing(Morales & Carolina, 2015). Ya no son los clásicos datos almacenados y estructurados en una base de datos. El universo del *Big Data* contempla la posibilidad de utilizar todos los datos disponibles a través de correos electrónicos, documentos, mensajes, imágenes, grabaciones de audio, registros, videos, incluso la transformación de datos de un formato a otro para que sean entendibles por otros sistemas(López García, 2013).

2.1.2 Tecnologías *Big Data*

Dentro de la arquitectura de un ambiente *Big Data* se pueden utilizar diferentes herramientas, cada una de estas cumple un papel importante. A continuación se realiza una descripción de las tecnologías más representativas que pueden implementarse.

***Hadoop*:** "Es un marco de trabajo - *framework* que permite el procesamiento distribuido de grandes conjuntos de datos a través de grupos de ordenadores que utilizan modelos de programación simple. Está diseñado para detectar y controlar los errores en la capa de aplicación"**(Guerrero López, Rodríguez Pinilla, & others, 2014)**. Ver la arquitectura de *Hadoop* en la figura 1.

Figura 1: Arquitectura Hadoop



Fuente: («Big data de código abierto para el impaciente, Parte 1», 2013)

***MapReduce*:** "El modelo de programación *MapReduce* se basa en dos funciones llamadas *Map* y *Reduce*. La entrada a dicho modelo es un conjunto de pares clave/valor y la salida es otro conjunto de pares clave/valor"(Goyzueta Rivera, 2015).

Función *Map*: A partir del conjunto de pares clave/valor de entrada se genera un conjunto de datos intermedios. La función *Map* asocia claves idénticas al mismo grupo de datos intermedios. Cada grupo de datos intermedios estará formado por una clave y un conjunto de valores por lo tanto, estos datos intermedios van a ser a su vez la entrada de la función de Reduce(Goyzueta Rivera, 2015).

Función Reduce: La fase de Reduce se encargará de manipular y combinar los datos provenientes de la fase anterior para producir a su vez un resultado formado por otro conjunto de claves/valores(Goyzueta Rivera, 2015).

Apache *Spark:* Es un *framework* de computación en paralelo que genera velocidades hasta 100 veces mayores que las desarrolladas por *Hadoop Mapreduce* en memoria relacionados con el procesamiento de datos a gran escala. Este *framework* es considerado la evolución de *Hadoop*. Es una alternativa interesante, especialmente en aplicaciones que requieren iteraciones y reusó de los datos (como el análisis de grafos y aprendizaje de máquina).

Machine learning: Tecnología aplicada a los datos que permite realizar predicciones basados en modelos generados del análisis de grandes cantidades de datos. “Es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos.”(Ferri, Ramírez, & Hernández, 2004)

Procesamiento de lenguaje natural: Es utilizado en la interpretación de sentimientos en especial en el social media. Según Chowdhury “Es un área de investigación que explora cómo las computadoras pueden utilizarse para entender y manipular texto escrito en lenguaje natural o del habla para hacer operaciones útiles”(Vivancos Vicente, 2016).

2.2 Cliente

“Cliente es la persona, empresa u organización que adquiere o compra de forma voluntaria productos o servicios que necesita o desea para sí mismo, para otra persona o para una empresa u organización; por lo cual, es el motivo principal por el que se crean, producen, fabrican y comercializan productos y servicios” (Arana & Hernan, 2015).

La definición anterior se ajusta en el contexto de medios impresos, partiendo del hecho que los clientes de este tipo de organizaciones son las personas o empresas que pautan avisos publicitarios, colocan avisos clasificados, suscriptores y público en general que adquiere cualquier servicio o producto que la organización ofrezca.

La famosa frase de Peter Drucker: “No se puede administrar, lo que no se puede medir” (Behn, 2005) fue muy sabia al predecir lo que está pasando en la actualidad, en particular con la información digital generada. Ahora con el *Big Data* y con distintas herramientas los gerentes pueden medir y saber precisamente de

una forma radical lo que está sucediendo en sus negocios y traducir directamente ese conocimiento en una toma de decisiones mejorada y en un rendimiento superior.

2.3 Estrategia

Estrategia es un término cuyo origen se remonta a la antigua Grecia y China, utilizado en torno a la guerra y en general a las acciones militares, posteriormente fue incluida en el lenguaje común de las organizaciones.

Una definición básica de estrategia sería: “Conjunto de acciones planificadas sistemáticamente en el tiempo que se llevan a cabo para lograr un determinado fin o misión” (Ardila Cañas & Gómez Díaz, 2015).

Desde la perspectiva empresarial una de las definiciones que más se ajusta define la estrategia como “La forma en la cual una corporación se comporta para diferenciarse a sí misma positivamente frente a sus competidores usando sus fortalezas relativas para satisfacer mejor las necesidades de los clientes” (García, 2010).

Uno de los sectores que quizás ha sido más influenciado por las tendencias tecnológicas son los medios de comunicación impreso. Los cuales están constantemente revisando sus diferentes estrategias entorno a mantenerse competitivos en el mercado y tener un mejor acercamiento con sus clientes.

Una de las ventajas que proporciona la implementación de Tecnologías *Big Data* repercute en la capacidad que adquiere la organización para procesar grandes volúmenes de datos, incluir nuevas fuentes de datos y aumentar la velocidad en obtener los datos a procesar y posteriormente disponer de información para generar conocimiento que permitirá tomar mejores decisiones y estrategias que permitan a la organización cumplir con sus objetivos.

Big Data supone un proceso de cambio en la organización, no sólo desde la perspectiva tecnológica sino principalmente desde la de negocio.

2.4 Segmentación

Desde la visión de mercadeo según Martínez y Milla (2012) (Silva Guerra, 2014). Afirman que la segmentación de mercado “se fundamenta en fragmentar el mercado total de un bien, en diferentes grupos más pequeños y semejantes que

comparten rasgos o atributos particulares". Para los periódicos la segmentación de mercado tiene las mismas bases expuestas en la definición anterior. Como a cualquier otra empresa, a un medio de comunicación impreso como en este caso, le interesa segmentar sus clientes para conocer sus características, necesidades e intereses particulares.

La promesa de valor de *Big Data* en este caso está representada en el perfeccionamiento del proceso de segmentación y fortalecimiento de las relaciones con los clientes. Esto se logra gracias al conocimiento de nuevas facetas de comportamientos del cliente que se conocen al incluir nuevas fuentes de datos no estructurados que antes eran imposibles de analizar debido a su naturaleza. De esta forma, se obtiene una visión más completa de las motivaciones de los clientes y se comprende mejor cuáles son sus preferencias.

2.5 Activos De Información

Hoy en día uno de los principales activos de las organizaciones es la información con la que se cuenta y el conocimiento de la misma. Esta se consolida a partir de datos de los clientes, proveedores, servicios, productos, competencia, etc. Pero ya no se trata solo de datos, las empresas normalmente tienen toda esta información en diferentes fuentes, por lo cual se requiere herramientas y estrategias que permitan extraerla y analizarla para obtener este importante activo para las empresas – el conocimiento (Acosta Medellín & Florez Lara, 2015)

Es importante aclarar el manejo que se le da a términos como lo son dato, información y conocimiento:

Datos. Son elementos que representan hechos relativos a un fenómeno o al resultado de un proceso en particular, los cuales carecen de significado por sí mismos ya que están fuera de un contexto que les dé sentido.(Silva Guerra, 2014).

Información. Debe entenderse como datos que han sido procesados, elaborados y además han sido situados en un contexto específico, por lo cual tienen un significado para alguien en un momento y un lugar determinados(Silva Guerra, 2014).

Conocimiento. Consiste en datos y/o información organizada y procesada para distribuir entendimiento, experiencia, aprendizaje acumulado y habilidades que pueden ser aplicados a un problema o actividad actual. El conocimiento en este

sentido es información que es contextual, relevante y sobre la que se puede actuar. Por tanto, la implicación es que el conocimiento tiene un fuerte componente de elementos de experiencia y reflexión que lo distingue de la información en un contexto determinado. Tener conocimiento implica que puede ser empleado para resolver un problema(Silva Guerra, 2014).

En el trabajo desarrollado las Tecnologías de *Big Data* representa una clara alternativa que permitirá incluir los diferentes activos de información con los que cuenta el periódico, para mejorar el conocimiento de los clientes, logrando así plantear modelos de análisis y aprovechar la información generada para la formulación de estrategias comerciales más exitosas.

Concluyendo podemos decir que las Tecnologías *Big Data*, son la herramienta imprescindible que deben utilizar los medios de comunicación en especial los periódicos para poder obtener ventajas competitivas que les permita generar valor para el negocio a partir de la toma de decisiones en base al análisis de los datos.

Big Data se convierte en un instrumento esencial de la estrategia comercial y de comunicación en cualquier tipo de organización, independiente de la actividad que desarrollen. Las organizaciones que disponen de mayores cantidades de datos permiten realizar mejores procesos de segmentación, lo cual se refleja en mejores estrategias comerciales.

Mayor cantidad de información acerca de los clientes representa mejores oportunidades para mejorar la “experiencia de cliente” lo cual se ve reflejado en mejores estrategias de fidelización.

Cada organización tiene una forma única de abordar proyectos basados en *Big Data*, esto se conoce como “Estrategias de adopción de tecnologías de *Big Data*”. Si bien es cierto que este tipo de Tecnologías son una muy buena práctica asociada a la ejecución de estrategias comerciales, no todas las soluciones son iguales, así que el paso siguiente es la definición de los elementos particulares para cada organización, de acuerdo a la medida de sus capacidades económicas, tecnológicas y de conocimiento sobre las misma por parte de su recurso humano. Este es sin duda el mayor aporte del presente proyecto y su dimensionamiento para el periódico Vanguardia Liberal.

3. DIAGNÓSTICO DEL ESTADO ACTUAL

En este capítulo se describe el estado actual de la empresa a nivel de los componentes tecnológicos que administran los procesos de negocio que se verán afectados por la definición del plan de implementación de tecnologías *Big Data*. De allí la importancia de describir los procesos de venta de las unidades de negocio de Publicidad, Circulación y Servicio al Cliente. Se hace una descripción de los sistemas de información que están involucrados en la administración de la información de cada una de las unidades de negocio y su forma de almacenamiento, luego se describen los flujos de venta de estas dependencias planteando una perspectiva de mejoramiento del proceso. Con este diagnóstico se evidencian las islas de información existentes y se hace una identificación inicial de los activos de información.

3.1 Arquitectura

Vanguardia Liberal dentro de su estructura organizacional cuenta con dos unidades de negocio que soportan toda su fuerza de ventas: Publicidad y Circulación, como unidad de soporte a sus operaciones cuenta con la unidad de negocio de Administración. A continuación mencionamos las características más importantes de cada uno de los componentes que conforman cada una de las unidades de negocio presentes en la arquitectura planteada en la figura 2.

3.1.1 Publicidad

Publicidad: Sistema comercial dedicado a gestionar el proceso de venta, pauta, publicación, facturación y administración de la cartera de los anunciantes del periódico, mantiene un control especial sobre la información relacionada con el cliente en lo referente a sus datos básicos, geográficos, comportamentales y transaccionales. Es un desarrollo propio en Visual FoxPro cuya tecnología de almacenamiento son tablas libres.

Estadístico: Desarrollo propio en Visual FoxPro, su función principal es ser repositorio de todas las ventas a crédito y/o contado generadas por las diferentes oficinas y agencias de la empresa, permite la generación de informes de tipo estadístico. Su tecnología de almacenamiento son tablas libres.

Punto de Venta: Controla lo relacionado con las ventas de contado o pagos relacionados con el ingreso de clasificados por palabra, pauta publicitaria, venta

de periódicos, retales, productos optativos entre otros que se presentan en los diferentes puntos de atención. Desarrollo propio en Visual FoxPro. Su tecnología de almacenamiento son tablas libres.

Sd-Class: Sistema comercial especializado en la venta de avisos clasificados por palabras, desarrollado por un tercero en Visual C++. Su tecnología de almacenamiento es soportada por Bases de datos relacionales SQL *Server*.

CRM: Gestión de recursos de clientes, Control de relacionamiento de clientes de publicidad, es un desarrollo propio en PHP, administra los clientes asignados a cada asesor y permite llevar un control de actividades realizadas. Su tecnología de almacenamiento es MYSQL. Actualmente está siendo rediseñado

Cotizaciones: Sistema encargado de realizar cotizaciones de trabajos comerciales, basado en una fórmula de costeo específica para la fabricación de periódicos a terceros. Sistema propio desarrollado en Visual FoxPro, su tecnología de almacenamiento son base de datos Visual FoxPro.

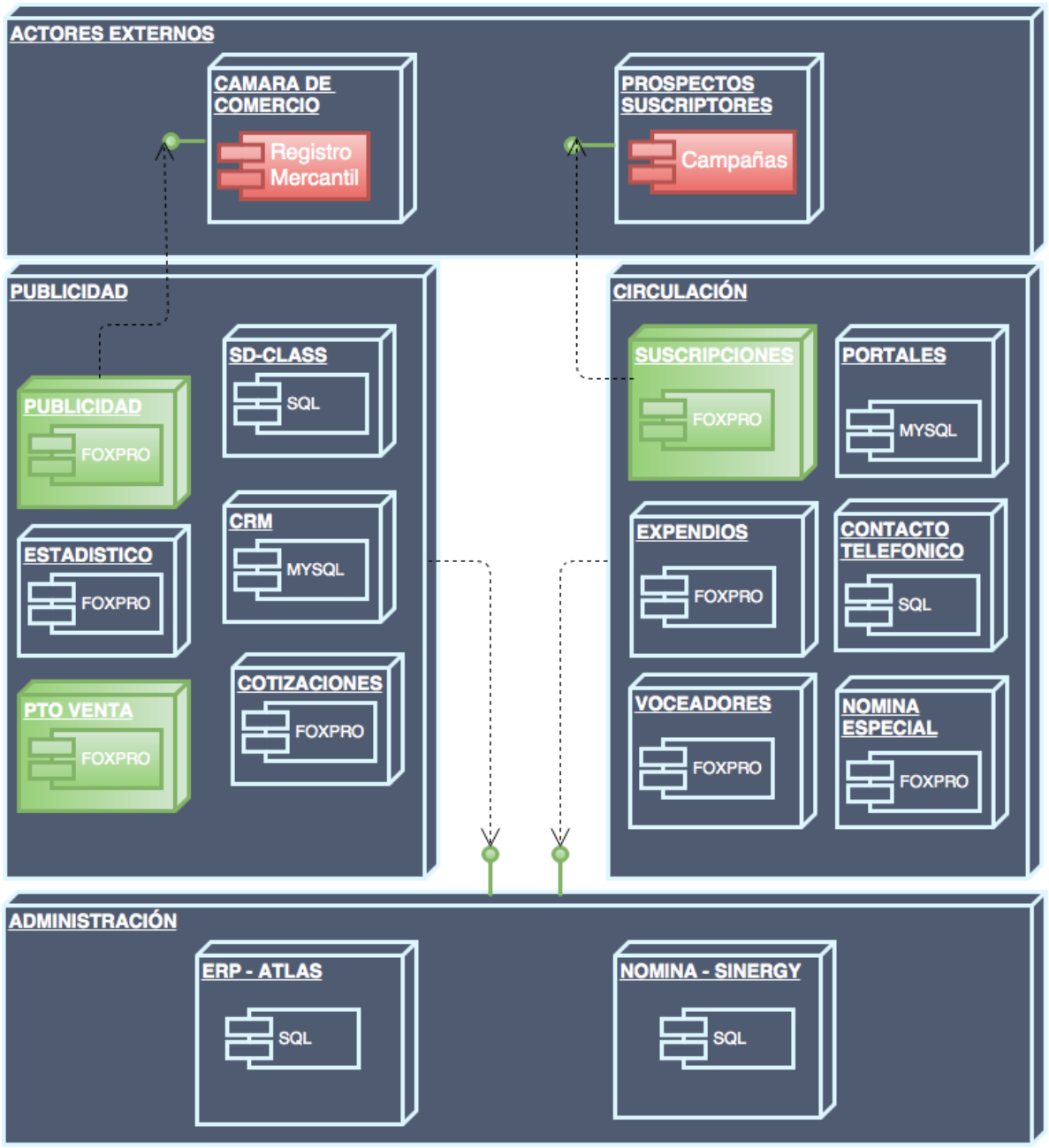
3.1.2 Circulación

Suscripciones: Sistema comercial encargado de administrar, gestionar y controlar el proceso de negocio que implica una suscripción al periódico Vanguardia Liberal, el cual incluye el manejo de la información del cliente en sus diferentes dimensiones tales como sus datos de identificación, descriptores de tipo geográfico, demográfico, comportamentales y de tipo transaccional. Además controla la logística que implica el reparto del periódico, los ciclos de facturación, el manejo de renovaciones, etc. Es un desarrollo propio en Visual FoxPro que cuenta con una forma de almacenamiento en tablas libres.

Expendios: Administra lo referente a distribución, venta y devolución de periódicos y productos optativos por parte de terceros con ubicación fija, dentro de sus procesos esta la facturación y control de reparto de los productos distribuidos. Está desarrollado en visual FoxPro y almacena la información en tablas libres.

Voceadores: Controla el despacho y la devolución de los terceros que ofrecen los periódicos y productos optativos de la empresa en la calle. Desarrollo propio en Visual FoxPro. Su información es almacenada en tablas libres.

Figura 2: Arquitectura Tecnológica



Fuente: Autor

Portales: Administran la información ingresada por las personas interesadas en las diferentes campañas promocionales ofrecidas por la empresa como concursos y sorteos. Es un desarrollo propio en PHP su información es almacenada en MYSQL.

Contacto Telefónico: Encargado de gestionar el proceso de renovación de suscripciones, registra las llamadas realizadas al cliente y genera programación de cobros. Desarrollo propio en .NET y su tecnología de almacenamiento es soportada por Bases de datos relacionales SQL *Server*.

Nómina Especial: Gestiona lo relacionado con la generación de comisiones y bonificaciones de la fuerza de venta de la unidad de negocio de circulación. Desarrollo propio en Visual FoxPro, su tecnología de almacenamiento son tablas libres.

3.1.3 Administración

ERP - Atlas: Sistema desarrollado por terceros en Visual FoxPro y .NET, su tecnología de almacenamiento es soportada por Bases de datos relacionales SQL *Server*, Administra lo relacionado con Contabilidad, Compras, Inventarios, Activos Fijos, entre otros módulos.

Nomina - *Sinergy*: Sistema de terceros especializado en la administración de nóminas, gestión de personal, pago de parafiscales, etc. Desarrollado en *Power Builder*, su tecnología de almacenamiento es soportada por Bases de datos relacionales SQL *Server*.

3.1.4 Actores Externos

Cámara de Comercio: Base de datos adquirida por la compañía cada 2 o 3 años con la información de las empresas grandes, medianas y pequeñas registradas en el registro mercantil de la cámara de comercio de Bucaramanga, con el fin de conocer información relevante para procesos de validación en el sistema de publicidad y generación de nuevos prospectos de anunciantes. Es entregada en formato de Excel y procesada para su normalización e inclusión dentro de los procesos de validación del sistema de publicidad.

Prospectos Suscriptores: bases de datos que se utilizan con el fin de contactar a los potenciales clientes de suscripciones, pueden ser producto de convenios con otras empresas, las cuales suministran información en Excel, archivos planos, etc. o cuponeras entregadas en los diferentes puntos de atención. Generalmente este tipo de información representa un proceso manual, en el caso de las cuponeras, o de normalización en el caso de la información digitalizada, para poder ser ingresada en los procesos de suscripciones.

Como se puede apreciar la arquitectura expuesta anteriormente está conformada por múltiples plataformas de almacenamiento, diferentes formatos de información, poca interoperabilidad entre los sistemas y necesidades comunes como la centralización de la información del cliente en un solo repositorio. Todas estas son el escenario común que soporta la adopción de tecnologías de *Big Data*.

3.2 Procesos De Venta

En este apartado revisaremos las características de los tres principales procesos de venta de la empresa de acuerdo al flujo de información en cada uno de ellos, también se propone para cada uno de ellos una perspectiva de mejora.

3.2.1 Proceso de venta de publicidad

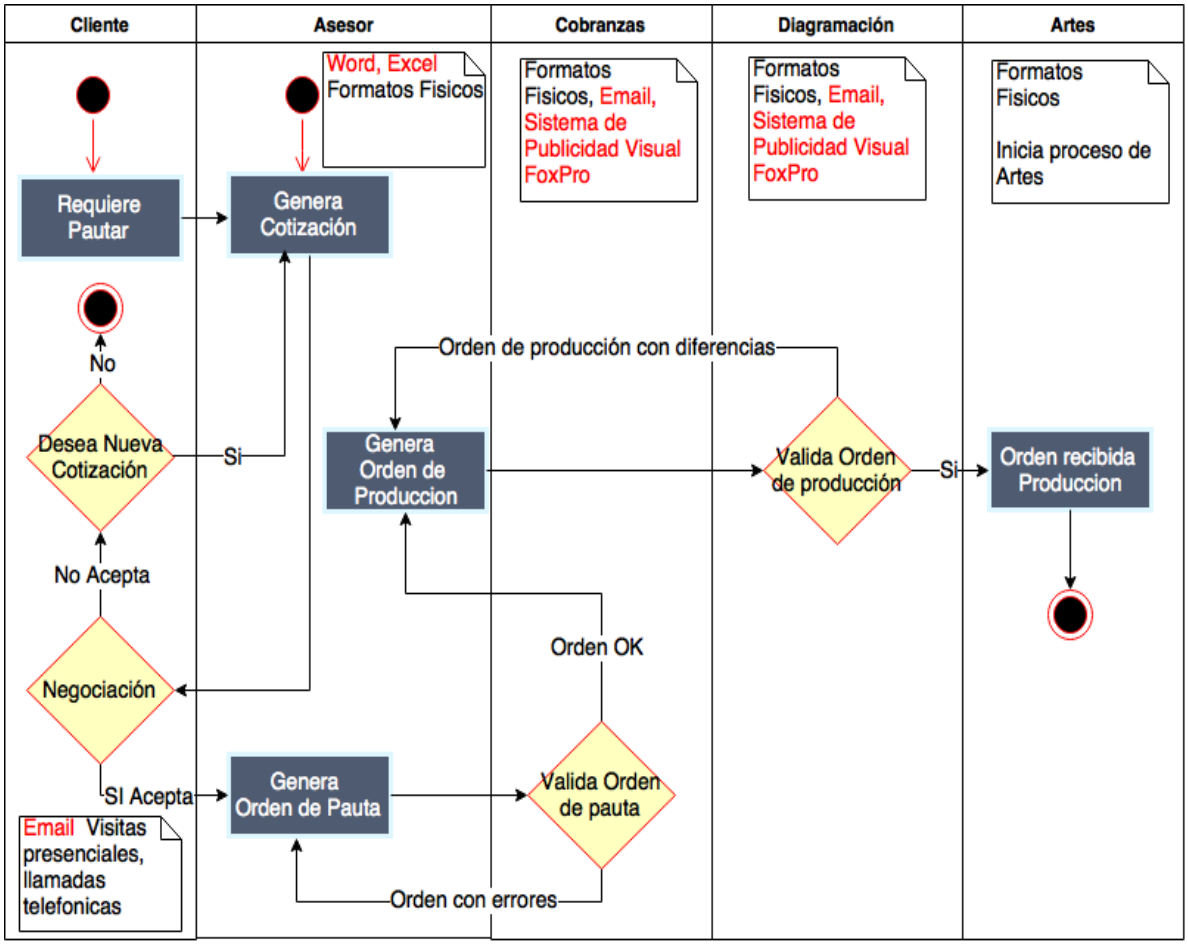
Tal y como se muestra en la figura 3, la venta de una pauta publicitaria puede iniciar desde el cliente o desde el asesor, en cualquiera de los dos casos el primer documento generado es una cotización, elaborada por el asesor en *Word* o *Excel*, esta orden pasa por un proceso de negociación hasta que el cliente acepta la cotización, en ese momento el asesor diligencia un formato pre impreso establecido denominado orden de pauta.

El asesor debe entregar o hacer llegar la orden de pauta firmada por cliente al área de cobranzas de la empresa, quien realiza un proceso de validación de la orden de pauta en cuanto a diligenciamiento de la misma, tarifas, descuentos y aprobaciones, si la orden presenta algún inconveniente es rechazada en cuyo caso el asesor realiza las correcciones e itera nuevamente con cobranzas hasta lograr la aprobación y la recepción de la orden por el área de cobranzas, a quien le corresponde ingresarla en el sistema de publicidad, este sistema se encarga del control de la publicación de los avisos pautados y de su correspondiente facturación y administración de cartera.

Cuando el asesor obtiene el visto bueno del área de cobranzas, procede con el diligenciamiento de la orden de producción y entrega de la misma al área de diagramación, quienes son los encargados de verificar que la orden de producción esté de acuerdo a las especificaciones plasmadas en la orden de pauta además de los datos específicos necesarios adicionales para la elaboración de los artes por parte de los diseñadores gráficos. Si Diagramación encuentra errores en la elaboración de la orden de producción la rechaza en cuyo caso el asesor realiza las correcciones e itera nuevamente con diagramación hasta lograr el visto bueno.

El área de Artes solo recibe las órdenes de producción con el visto bueno del área de diagramación. Esta área es la encargada de elaborar los artes de los avisos y validarlos con el asesor para su publicación.

Figura 3: Diagrama de proceso de venta publicidad



Fuente: Autor

Perspectiva de mejoramiento tecnológico del proceso

El proceso de venta actualmente está sistematizado desde el ingreso de la orden de pauta, pero todo lo que implica la negociación entre el cliente y el asesor es desconocido por la empresa, aspectos como: Cuantas cotizaciones fueron necesarias para llegar a una negociación en firme, Cuál fue el valor inicial cotizado versus el negociado, cuantas llamadas, cuantas cotizaciones, correos visitas presenciales fueron necesarias para concretar una negociación, son datos importantes que hoy se desconocen.

Involucrar al asesor en el ingreso de sus propias órdenes de pauta al sistema, reduciría sustancialmente los errores que se presentan actualmente por la transcripción de los diferentes formatos físicos utilizados durante el proceso de venta. El tiempo de los asesores seria optimizado debido a que sus desplazamientos para entregar las órdenes diligenciadas ante el auxiliar del área de cobranzas serían innecesarios, los flujos de autorización serán automatizados, de tal forma que el asesor no perderá tiempo buscando autorizaciones para las negociaciones realizadas ya que estas serán solucionadas o rechazadas por su superior a través de un flujo de autorización establecido. Centralizar la información de los clientes permitirá que las diferentes áreas comerciales de la empresa conozcan la misma información y comportamiento en los demás canales de venta de la organización.

3.2.2 Proceso de Venta en servicio al Cliente

En la figura 4 se puede apreciar el proceso de la venta de contado de periódicos, avisos publicitarios, productos optativos y demás en los puntos de atención al cliente, inicia cuando el cliente hace la solicitud del producto o manifiesta que desea hacer el pago de facturas de suscripciones, Expendios o clasificados. La cajera realiza entonces un proceso de validación el cual puede consistir en verificar si hay inventario disponible del producto solicitado, también si es el caso verifica el valor a pagar de la última factura generada por el sistema de suscripciones o expendios, puede también verificar la existencia de la factura de clasificados que el cliente desea cancelar.

Paso siguiente la cajera solicita los datos básicos o confirma los mismos si el cliente ya figura como registrado en la base de datos del sistema de punto de venta. Después se genera la factura o recibo soporte del pago realizado, si es el caso del pago de una suscripción o factura de Expendios. Por último se procede con la entrega de la factura o soporte de pago y producto al cliente.

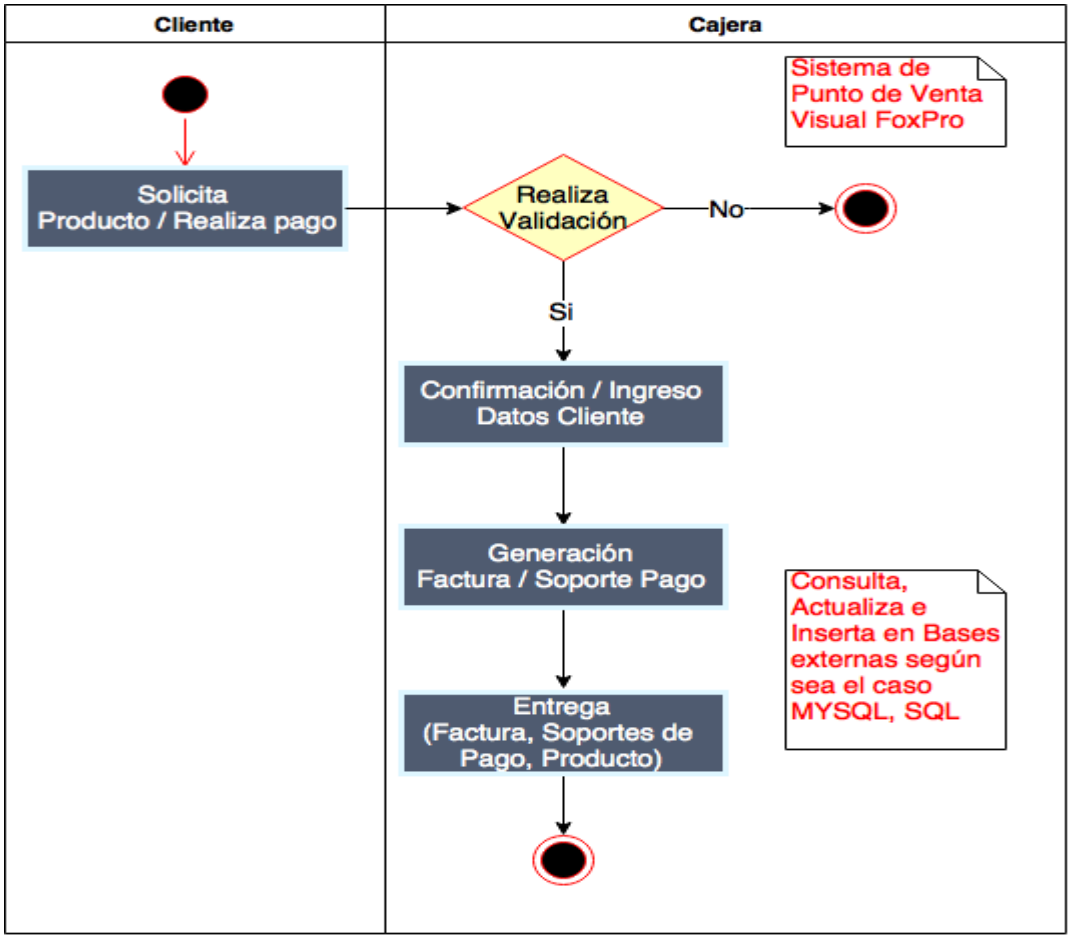
Perspectiva de mejoramiento tecnológico del proceso

Reducir el riesgo de ingresar información errada de los clientes que compran de contado al poder validar la información contra fuentes externas de forma ágil, sería un aspecto importante a implementar.

Disponer de la información centralizada permitirá mejorar la experiencia de cliente, ya que la persona de caja podrá conocer si el cliente atendido es suscriptor y por consiguiente tiene precios preferenciales en ciertos productos. Conocer la información de tipo transaccional le puede llegar a indicar al cliente productos

complementarios con los ya adquiridos en otras ocasiones, de esta misma forma la persona encargada de atender al cliente podrá saber si el cliente ya está inscrito en los diferentes concursos o promociones en los cuales puede llegar a participar, todo este tipo de información que el usuario puede llegar a tener disponible en ese momento repercutirá en la fidelización de mayor cantidad de clientes.

Figura 4: Diagrama de proceso de venta punto de venta



Fuente: Autor

3.2.3 Proceso de venta suscripciones

La venta de suscripciones es posible que se presente desde una iniciativa propia del cliente, otra forma puede ser por gestión del asesor de ventas que logra vender contratos nuevos de suscripción y finalmente el canal de renovaciones logra que los suscriptores renueven su suscripción por un periodo más. Ver figura 5.

Un contrato nuevo inicia con una propuesta de opciones de suscripción que el asesor ofrece, si el cliente acepta la propuesta, se procede a la elaboración del contrato físico. El contrato es presentado ante el auxiliar de suscripciones, quien es el encargado de ingresarlo al sistema de suscripciones, en donde se valida la información del mismo y se programa el inicio de la suscripción. En caso de no aceptar la propuesta por parte del asesor finaliza el proceso.

En el caso de renovaciones se contacta al cliente y se le pregunta por su intención de renovar la suscripción por un periodo más, si acepta se pacta una fecha de pago y se continúa con la suscripción. En caso de no aceptar la renovación se termina el proceso.

Perspectiva de mejoramiento tecnológico del proceso

Los aspectos a mejorar en el canal de venta de suscripciones son muy similares a los del canal de venta de publicidad tales como: La eliminación de formatos físicos, los desplazamientos innecesarios de los asesores, la reducción de errores en la transcripción del formato físico al sistema y el involucramiento del asesor en el proceso de ingreso de la información al sistema.

Es muy importante minimizar en lo posible los riesgos de ingreso de información errada mediante la utilización de fuentes externas de validación lo anterior desde la perspectiva de suscriptores nuevos. Otro aspecto relevante en el canal de suscripciones es la fidelización de los clientes a través de la renovación de su suscripción. En este punto es fundamental conocer qué tipo de cliente es, cuántas llamadas fueron necesarias hacer para lograr su renovación en el periodo anterior, qué tipo de beneficio se le otorgó, etc.

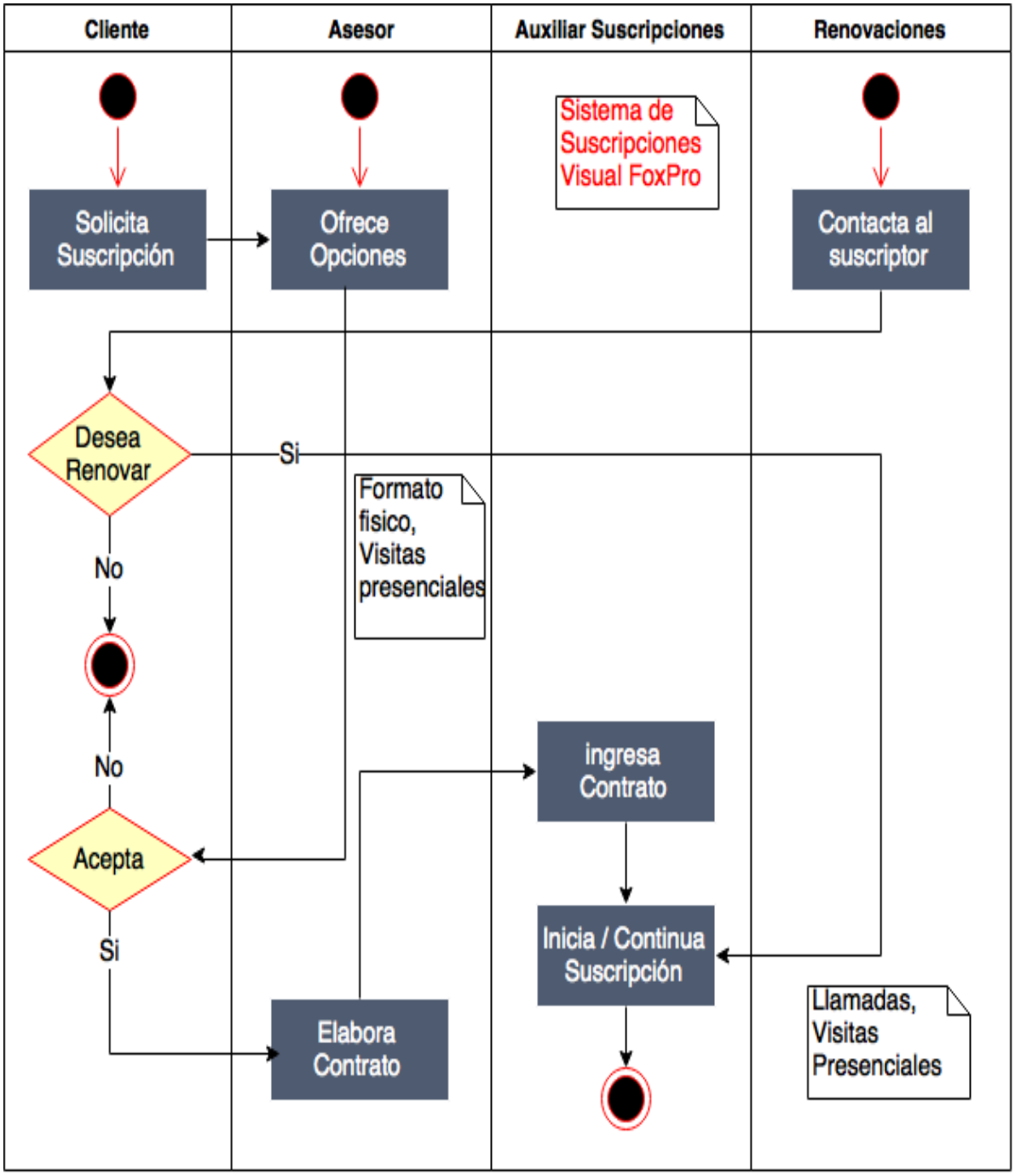
El conocimiento del cliente en especial en este canal se convierte en un verdadero tesoro que puede llegar a mejorar considerablemente las ventas de productos optativos a nivel de suscriptores.

3.2.4 Proceso de Validación de la información de los Actores externos

Dentro del proceso de validación de Actores externos de información se lleva a cabo el análisis de la fuente externa en su estado natural, luego se procede a transformar en un formato que facilite el tratamiento de los datos, FoxPro, seguidamente se realizan tareas de eliminación de duplicados, registros incompletos, estructuración de la información, normalización, estandarización y

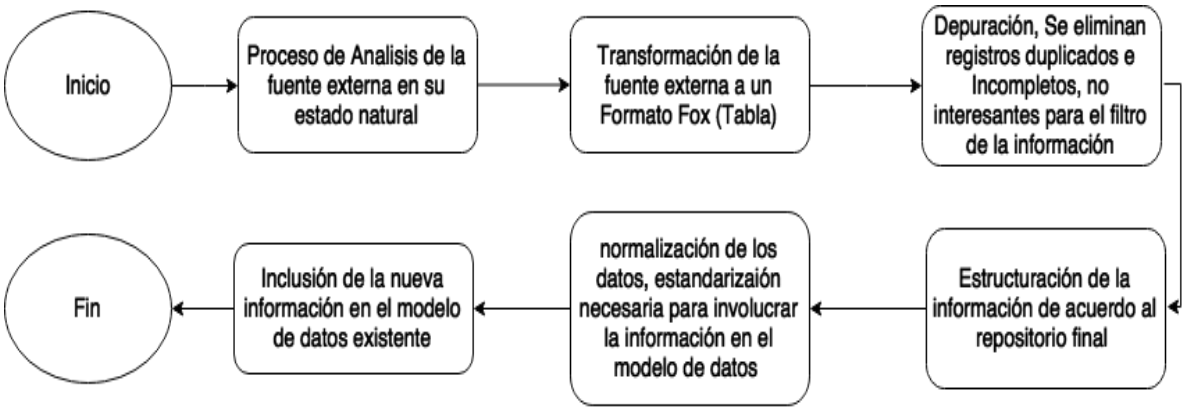
formateo de los datos, para finalmente incluir esos actores externos dentro del modelo de datos. Ver figura 6.

Figura 5: Diagrama de proceso de venta suscripciones



Fuente: Autor

Figura 6: Proceso de validación actores externos



Fuente: Autor

Perspectiva de mejoramiento tecnológico del proceso

El proceso como tal puede llegar a ser mejorado en aspectos de velocidad en lo que tiene que ver con la transformación estructuración, normalización y formateo de la información. Ya que actualmente estas tareas son manuales.

4. IDENTIFICACIÓN DE REQUERIMIENTOS

En este apartado se da a conocer el desarrollo de la encuesta realizada a Gerentes, Sub Gerentes y Coordinadores de área de las unidades de negocio de publicidad, circulación, administración y tecnología. Cuyo objeto era identificar las necesidades de gestión de información, principales variables de segmentación y las fuentes de información externa e interna de interés dentro de sus procesos de negocio.

4.1 Encuesta

La metodología planteada contempla la identificación de activos de información, para adelantar este proceso y después de presentar los procesos actuales se procede a aplicar un instrumento de consulta a los conocedores de las unidades de negocio. El instrumento busca identificar los principales atributos y características que permitan una mejor formulación de estrategias comerciales y de segmentación en Vanguardia Liberal.

4.1.1 Instrumento de recolección de la información

Para determinar los atributos y características más importantes en el modelaje de planteamientos de estrategias comerciales y de segmentación en Vanguardia Liberal, se utilizó como instrumento de recolección de la información la encuesta; Este instrumento es de gran importancia ya que permite conocer la opinión en relación con los objetivos de la investigación de un grupo de personas seleccionadas. Según Trespalacios, Vázquez y Bello (Gutiérrez, Acebrón, & Casielles, 2005) “La encuesta es utilizada en investigaciones de tipo descriptiva, desde la cual se precisan identificar las preguntas a realizar, las personas seleccionadas en una muestra representativa de la población, especificar las respuestas y determinar el método empleado para recoger la información que se vaya obteniendo”. Dentro de la investigación se empleó la encuesta ya que por medio de esta técnica se logró conocer de forma precisa la opinión de los encargados de la gestión de información de los tres procesos de venta identificados: Publicidad, Circulación y Servicio al Cliente. En la tabla 6, se detalla la ficha técnica de la encuesta.

En el anexo A, se puede ver la encuesta realizada a catorce personas encargadas de la gestión de la información en Vanguardia Liberal: Gerentes, Sub-Gerentes y

Coordinadores de área de las unidades de negocio de Publicidad, Circulación, Administración y Tecnología.

La encuesta tuvo como propósito realizar una estimación de tipo cualitativo, donde se emplea el método de estudio de caso con el fin de plantear tecnologías de *Big Data* para optimizar las estrategias comerciales y de segmentación de Vanguardia Liberal, el objetivo de este instrumento de recolección de información fue conocer la opinión de los encargados de gestionar la información de Vanguardia Liberal, con el fin de recibir sus contribuciones en relación a la gestión de grandes volúmenes de información estructurada y desestructurada para la implementación de soluciones y/o proyectos de *Big Data*. Esta encuesta fue creada a través de la herramienta *Google Docs*. En el numeral 4.3, se presenta el análisis de los resultados producto de la encuesta realizada.

Basados en la definición de los objetivos se buscó orientar la encuesta para identificar aquellas fuentes de información relevantes tanto externas como internas que cubrieran las necesidades de gestión de los procesos de negocio y la identificación de las variables más relevantes para procesos de segmentación.

Tabla 6: Ficha Técnica

FICHA TÉCNICA	
Persona quien realizó la encuesta:	Francisco Carrillo Álvarez
Grupo Objetivo:	Gerentes, Sub Gerentes y Coordinadores de las unidades de negocio de Publicidad, Circulación y Administración
Tamaño de la Encuesta:	14 encuestas reales, 14 encuestas ponderadas
Técnica de Recolección de Datos:	Diligenciamiento de formulario desarrollado y publicado en *Google Docs*
Tipo de la Muestra:	La muestra es dirigida, los encuestados seleccionados son empleados de Vanguardia Liberal

FICHA TÉCNICA	
Preguntas que se Formularon:	Ver anexo A
Tema o Temas a los que se Refiere:	Identificación de variables de segmentación, fuentes de datos a consultar e importancia de la analítica de datos masivos
Fecha de Realización:	Del 31/03/2016 al 05/04/2016
Área / Cubrimiento:	Vanguardia Liberal

Fuente: Autor

4.2 Resultados

Para revisar en forma detallada los resultados de la encuesta aplicada por favor remítase al Anexo B.

4.3 Análisis de los resultados

Primera pregunta: ¿Qué tan Importante considera los siguientes aspectos para segmentar su negocio? Al contestar tenga en cuenta la siguiente escala: 1. Nada Importante, 2. Un Poco Importante, 3. Muy Importante, 4. Extremadamente Importante. Para ver el listado de opciones, por favor remítase al Anexo 2.

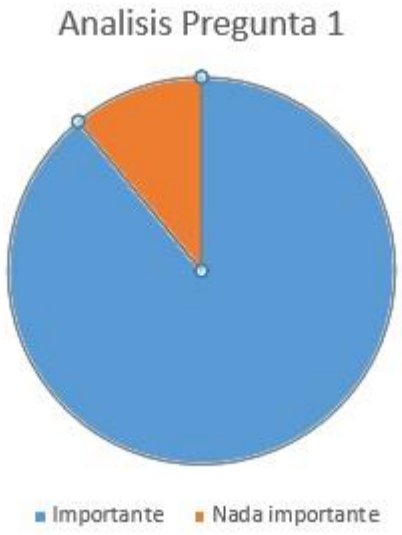
Con este planteamiento se busca determinar la importancia que representan las variables propuestas para realizar procesos de segmentación de clientes, revisando los resultados se puede observar que el 85% de los encuestados confirman en términos generales la importancia y necesidad de hacer procesos de segmentación con las características expuestas es importante. Ver figura 7.

Los resultados reflejan una clara tendencia por las características de segmentación de tipo demográfico como: Ubicación Geográfica, Fecha de Nacimiento, Género, entre las más importantes, seguidas de las de conducta como: Frecuencia de Compra, Monto compras último mes, Monto de compras último año, entre otras y finalmente las tipo psicográficas definidas por variables como *Facebook, Twitter, Instagram* y *Linkedin*. Ver Anexos 7, 8, 9, 10,11

Segunda pregunta: ¿Cuáles de las siguientes fuentes externas de información consideraría importantes incluir dentro de su proceso de negocio? Marque con una (X). Para ver el listado de opciones, por favor remítase al Anexo 3.

Aquí se identifican las fuentes de datos externas más apropiadas, así: La cámara de comercio de Bucaramanga, el RUT (Registro único tributario) y el RUES (Registro único empresarial y social), son consideradas por los encuestadas como las más importantes, seguidas de: El RUNT (Registro único de tránsito), el FOSYGA (Fondo de solidaridad y garantía), el RUAF (Registro único de afiliados), Aliados comerciales y otras fuentes propuestas por el grupo objetivo de la encuesta como: Las centrales de riesgo y las bases de datos de concursos ofrecidos por los diferentes portales web de la organización. Ver Anexo 12

Figura 7: Análisis Pregunta 1



Fuente: Autor

Tercera pregunta: Califique de acuerdo al nivel de relevancia la información interna con la cual le gustaría contar en la gestión de su proceso de negocio. Al contestar tenga en cuenta la siguiente escala: 1. Nada relevante, 2. Un poco relevante, 3 Muy Relevante, 4 Extremadamente Relevante. Para ver el listado de opciones, por favor remítase al Anexo 4

El 99% de los encuestados están de acuerdo en que todos los aspectos presentados son considerados Relevantes. Los resultados indican claramente una preferencia por conocer qué productos tiene activos el cliente con la empresa esto

es evidente en las variables de: Es cliente de pauta publicitaria, Es suscriptor del periódico, Es cliente de optativos, Es cliente de clasificados. De igual forma se percibe el interés en conocer en todo momento la inversión del cliente en cada uno de los canales de venta de la empresa esto se puede apreciar en los resultados de los aspectos como: Monto pauta publicitaria último periodo, Monto de optativos último periodo, Monto de clasificados último periodo. Ver Anexo 13.

Cuarta pregunta: ¿Qué nivel de importancia tendría para Ud., los siguientes aspectos para obtener una analítica de datos masivos que brinde un mejor apoyo en la gestión de su proceso? Al contestar tenga en cuenta la siguiente escala: 1. Nada Importante, 2. Un poco importante, 3. Muy Importante, 4. Extremadamente Importante. Para ver el listado de opciones, por favor remítase al Anexo 5.

En este punto se presentan aspectos referentes con la analítica de datos masivos, se busca definir el nivel de importancia de la inclusión de estos, como parte fundamental del proceso de gestión de su negocio. La evaluación de los resultados evidencia una clara alineación que se confirma con la alta importancia para todos y cada uno de los puntos enunciados. Ver Anexo 14.

Quinta pregunta: Basado en su interés y de acuerdo al proceso de negocio que gestiona, determine el nivel de impacto de los siguientes beneficios empresariales derivados del uso y análisis masivo de mayores fuentes de datos internas y externas. Al contestar tenga en cuenta la siguiente escala: 1. Nada Impactante, 2. Un poco impactante 3. Muy Impactante, 4. Extremadamente Impactante. Para ver el listado de opciones, por favor remítase al Anexo 6

En esta pregunta se pretende medir el nivel de impacto que percibe el grupo objetivo de la encuesta de acuerdo a los beneficios que puede llegar a tener la implementación de procesos de analítica de datos masivos en la gestión de su proceso de negocio. El 100% de los encuestados consideran que los beneficios presentados son extremadamente impactantes. Ver Anexo 15.

4.4 Conclusiones de los Resultados

En la pregunta 1 donde el objetivo principal era identificar las variables de segmentación más apropiadas, se presentó un fenómeno contradictorio entre la variable más representativa la Ubicación geográfica y el Código Postal, debido a que estas dos variables tienen una estrecha relación, lo cual puede explicarse por el desconocimiento del potencial de utilización que puede llegar a tener el código

postal a futuro en Colombia. Se determina que los resultados que indican que no es importante el Código postal, se deben a que el concepto de código postal en Colombia es relativamente nuevo. Se opta entonces tomarla como una variable significativa para el estudio.

Las dos fuentes externas propuestas por los encuestados, proponen la inclusión de las bases de datos de concursos y promociones que realiza el periódico, se acoge esta propuesta ya que estas bases serían un excelente aporte a procesos de validación de la información. Otra fuente propuesta fueron las consultas a las centrales de riesgo CIFIN, en este caso se toma la decisión de no tenerla en cuenta como fuente viable por la sensibilidad de la información que se maneja y los requerimientos exigidos por el ente controlador.

Se identificó el inventario de activos de información que serán objeto de tratamiento de los procesos que se establezcan: Publicidad, Suscripciones, Punto de Venta, Información de la Cámara de comercio, Información de campañas de prospectos de suscriptores.

5. TECNOLOGIAS BIG DATA

Esta sección del documento representa el proceso de selección y evaluación de las herramientas más convenientes a ser implementadas de acuerdo a las necesidades de Vanguardia Liberal. Se inicia con la presentación del ecosistema actual de herramientas de *Big Data*, Según Matt Turk presentado en Febrero de 2016(«*Is Big Data Still a Thing*?», 2016), luego se describen, comparan y se selecciona el tipo de solución que puede llegar a implementar el periódico. Seguidamente se revisan algunos reportes de los cuadrantes mágicos presentados en Febrero de 2016 por la firma consultora Gartner(«Acerca de Gartner», s. f.) y se procede a evaluar la distribución más representativa de la misma forma se toma la decisión sobre una herramienta de *Bussines Intelligence*. Finalmente se hacen unas consideraciones técnicas sobre las herramientas seleccionadas.

Existe todo un ecosistema de herramientas que pueden ser implementadas dependiendo de la necesidad asociada a *Big Data*. Nombres como *Hadoop*(Guerrero López et al., 2014), NoSQL(«Las bases de datos NoSQL», s. f.), Cassandra(«El Proyecto Apache Cassandra», s. f.), *Business Intelligence*(«5 ventajas de la Inteligencia de Negocios», s. f.), *Machine Learning*(Ferri et al., 2004), *MapReduce*(Goyzueta Rivera, 2015)… son solo algunos de los más conocidos. En este capítulo mencionaremos las tecnologías y herramientas más destacables y que mejor relación tienen frente a las temáticas a desarrollar en el presente trabajo. Partimos del panorama actual del *Big Data* publicado en febrero de 2016 por Matt Turk.(«Is *Big Data Still a Thing*?», 2016) (Ver Figura 8).

5.1 Principales soluciones tecnológicas de *Big Data*

En la tabla 7 se puede apreciar el listado de herramientas *Big Data* que tienen mayor afinidad y viabilidad de adopción teniendo en cuenta aspectos como: respaldo ofrecido a nivel de soporte por su casa matriz, evolución tecnológica, utilidades ofrecidas, implementación del ecosistema *Hadoop* dentro de sus componentes, facilidad de uso, escalabilidad, curva de aprendizaje, costo de licenciamiento, cantidad de socios de negocio ubicados en Colombia entre otros. En el Anexo C. Se describen cada una de las herramientas mencionadas en la tabla 7.

Figura 8: Panorama Big Data 2016

Fuente: («*Is Big Data Still a Thing*?», 2016)

Tabla 7: Herramientas Big Data

Herramienta	Proveedor	Tipo Solución	Enfoque
Soluciones IBM para *Big Data*	*International Business Machines Corp.*	*Appliance, Cloud*	Soluciones Analíticas
HP *Haven*	*Hewlett-Packard*	*Appliance, Cloud*	Soluciones Analíticas
Teradata	Teradata *Corporation*	Herramienta	Almacenamiento y Analítica
Oracle Fast Data Solutions	*Oracle Corporation*	Distribución	Transformación, Integración, Analítica, Tiempo real
SAP HANA	SAP	*Appliance*	Tiempo real
Soluciones Amazon	Amazon, Inc	*Cloud*	Almacenamiento, Procesamiento, Escalabilidad
Soluciones *Microsoft*	*Microsoft Corporation*	*Cloud*	Almacenamiento, Procesamiento, Escalabilidad
Soluciones *Google*	*Alphabet Inc.*	*Cloud,*	Almacenamiento, Procesamiento, *Machine Learning,* Escalabilidad
VMware	*VMware Inc*	Herramienta	Virtualización, Almacenamiento, Procesamiento
Cloudera	Cloudera	Distribución	Conectividad, Transformación, Integración, Almacenamiento, Procesamiento, Analítica
CloudAnt	*International Business Machines Corp.*	*Cloud*	Almacenamiento, Procesamiento, Escalabilidad

Herramienta	Proveedor	Tipo Solución	Enfoque
HortonWorks	*Hortonworks Inc.*	Distribución	Conectividad, Transformación, Integración, Almacenamiento, Procesamiento, Analítica
Splunk	*Splunk Inc.*	Herramienta	Monitoreo
MongoDB	MongoDB Inc.	Herramienta	Almacenamiento NoSQL
Cassandra	*Apache Software Foundation*	Herramienta	Almacenamiento NoSQL
Soluciones Apache *Hadoop*	*Apache Software Foundation*	Herramientas	Conectividad, Transformación, Integración, Almacenamiento, Procesamiento, Analítica
MapR Technologies	*MapR Technologies, Inc.*	Distribución	Conectividad, Transformación, Integración, Almacenamiento, Procesamiento, Analítica
Soluciones Pentaho	Pentaho Corporation	Herramientas	Conectividad, Transformación, Integración, Almacenamiento, Procesamiento, Analítica, Minería de datos y visualización
Soluciones Sap	SAP	Herramientas	Conectividad, Integración, Analítica
Proyecto R	*R Foundation.*	Herramienta	Programación, Analisis Estadístico y visualización

Herramienta	Proveedor	Tipo Solución	Enfoque
QlikView	*QlikTech International*	Herramienta	Analítica, Conectividad, tiempo real, Visualización
Pivotal	*Pivotal Software, Inc*	Distribución	Conectividad, Transformación, Integración, Almacenamiento, Procesamiento, Analítica
SPSS	*International Business Machines Corp.*	Herramienta	Análisis Estadístico, Visualización
Tableau	*Tableau Software, Inc*	Herramienta	Analítica, Conectividad, tiempo real, Visualización
MLlib	*Apache Software Foundation*	Herramienta	Machine Learning, Estadística
Weka	*The University of Waikato*	Herramienta	Analítica, Minería de Datos
Microsoft Power BI	*Microsoft Corporation*	Herramienta	Analítica, Conectividad, tiempo real, Visualización
Tibco SpotFire	*TIBCO Software Inc*	Herramienta	Analítica, Conectividad, tiempo real, Visualización

Fuente: Autor

5.2 Tipos de Soluciones Big Data

Para implementar un proyecto de Big Data, existen básicamente tres tipos de soluciones: Distribuciones*, Appliance* y *Cloud*, entre las cuales las organizaciones

deben escoger la más conveniente en cada caso. A continuación se describen cada una de las soluciones y se hace una comparación para tomar la decisión más adecuada para implementar en el periódico.(Sabater Picañol, 2013),(Serrat Morros, 2013)

5.2.1 Distribuciones

Las Distribuciones de Software, son básicamente *Clusteres* configurados según las necesidades particulares. Es decir que la solución se da mediante la instalación y configuración de las herramientas o tecnologías más apropiadas según el criterio de cada organización.

"Este tipo de solución es el ideal para comenzar un entrenamiento con herramientas *Big Data*, pues hay una gran cantidad de soluciones software *open source*".(Sabater Picañol, 2013) La principal ventaja de este tipo de soluciones es la flexibilidad a la hora de elegir las especificaciones para que estén dentro de unos límites presupuestarios y que a la vez cumplan con los requisitos del proyecto.

"Requieren de un conocimiento mínimo en administración de sistemas tanto para encontrar las especificaciones adecuadas como para el mantenimiento de las máquinas. Se requiere un conocimiento previo en distribuciones *Hadoop* para poder realizar la instalación y configuración del software".(Serrat Morros, 2013)

5.2.2 *Appliance*

Son productos completos ya instalados, configurados y optimizados que involucran hardware y software, de altas especificaciones técnicas. Estas infraestructuras son ofrecidas por compañías que se encargan de su instalación y servicio técnico. Su implementación tiene un costo elevado. Por ser un hardware muy específico que normalmente solo conocen a la perfección la empresa que lo distribuye, su mantenimiento y soporte son costosos.

Normalmente las compañías que venden *appliance Big Data* suelen vender módulos de ampliación que permiten escalar de forma horizontal en rendimiento y almacenamiento, no obstante no es tan práctico como comprar un nuevo servidor y añadirlo (como ocurre con las distribuciones). Las *appliance* permiten reducir el espacio que ocupan los clústeres, al estar los diferentes nodos del *clúster* encajados en un mismo módulo.(Sabater Picañol, 2013),(Serrat Morros, 2013).

5.2.3 *Cloud*

Son soluciones *Big Data* ofrecidas como servicios por compañías como Amazon, Microsoft, IBM entre otras. Sus principales características son:

- Se paga un costo de alquiler de infraestructura. (Se paga lo que se usa)
- Alta escalabilidad, puede crecer el número de nodos de forma sencilla.
- Alta Disponibilidad

“Las soluciones *cloud* están indicadas especialmente en casos en los que el análisis de información sea muy puntual (necesitas analizar varios terabytes de datos, pero sólo lo harás una vez), en caso de uso continuo saldría bastante costoso.”(Sabater Picañol, 2013)

“Dependiendo del tipo de información que se analiza puede ser sensible o puede no ser legal subir esos datos a un servicio externo a la empresa, dependiendo de la ley de protección de datos vigente del país.” (Sabater Picañol, 2013)

5.2.4 Comparación de Soluciones

Los parámetros utilizados para la valoración del tipo de solución *Big Data* más favorable varían dependiendo de las necesidades de cada empresa y del tipo de proyecto a implementar. Para el caso de estudio, las variables seleccionadas fueron planteadas con la expectativa de implementación del plan de tecnologías *Big Data* para la optimización de estrategias comerciales y de segmentación: facilidad de instalación, costos de mantenimiento, costos de implementación, escalabilidad, flexibilidad, rendimiento y entrenamiento.

Vanguardia Liberal en sus procesos de compra de tecnología a nivel interno considera siempre el costo de implementación, el costo del mantenimiento, el proceso de implementación (Instalación, Configuración. Acompañamiento del consultor y puesta a punto) entre otros. Como parámetros adicionales en esta evaluación se incluyen la escalabilidad, flexibilidad, rendimiento y experimentación como factores propios a evaluar en herramientas *Big Data*(Garcés Uquillas, 2015) según Consultoras de renombre como Gartner(«Acerca de Gartner», s. f.), Forrester(«Forrester : Welcome», s. f.), Dresner(«*Dresner Advisory Services - Home of Business Intelligence and the Wisdom of Crowds ® Market Research*», s. f.) y BARC(«The BI Survey 15, BARC’s annual repo*rt on the BI industry*», s. f.).

En el Anexo D se describen en detalle la escala utilizada para la valoración de cada uno de los aspectos evaluados. En la tabla 8 puede apreciarse el proceso comparativo realizado sobre los tres tipos de solución Big Data:

Costo de implementación: Las soluciones de tipo *appliance* son más costosas pero ofrecen simplicidad y rendimiento. En el caso de las soluciones *cloud* se paga un arriendo por el uso que se les da, haciéndolas más económicas en usos puntuales, son simples en su configuración y además no tienen que adquirir infraestructura. Las distribuciones son una solución intermedia, permite hacer la configuración de hardware y de software que la necesidad amerite, son ajustables a los presupuestos, se necesita cierto grado de conocimiento.(«*Big Data Appliance* | Oracle España», s. f.),(Cloudera, comerciales, & Aquí, s. f.),(«AWS | Análisis de *Big Data* y almacenamiento en la nube», s. f.)

Tabla 8: Características de Tipos de Soluciones Big Data

Características	**Distribuciones**	***Appliance***	***Cloud***
Costo de Implementación	3	1	3
Costo de Mantenimiento	3	1	3
Instalación y Configuración	1	3	2
Escalabilidad	2	1	3
Flexibilidad	3	2	1
Rendimiento	2	3	3
Experimentación	3	1	1
Total	**17**	**12**	**16**

Fuente: Autor

Costo de mantenimiento: Las soluciones de tipo *appliance* necesitan de un experto que conozca el software y hardware su mantenimiento y soporte son más costosos. Las compañías que ofrecen servicios de infraestructura en la nube se ocupan del mantenimiento hace que los costos por este parámetro disminuyan considerablemente. Se necesita de personal calificado que solucione los inconvenientes pero su costo es mucho más económico que las *appliance*.(«*Big*

Data Appliance | Oracle España», s. f.),(Cloudera, comerciales, et al., s. f.),(«AWS | Análisis de *Big Data* y almacenamiento en la nube», s. f.)

Instalación y configuración: En el caso de las soluciones basadas en distribuciones es indispensable conocimiento a nivel de administrador, para decidir la configuración apropiada. *Las appliances* están configuradas para un rendimiento óptimo y las *cloud* se configuran a necesidad, simplificando este proceso.(«*Big Data Appliance* | Oracle España», s. f.),(Cloudera, comerciales, et al., s. f.),(«AWS | Análisis de *Big Data* y almacenamiento en la nube», s. f.)

Escalabilidad: En los ambientes de tipos de solución *appliances* y distribuciones escalar en capacidad de procesamiento y almacenamiento implica adquirir nuevas máquinas. En el caso de las *appliances* se debe adquirir el hardware específico de cada *appliance*. Para *cloud* simplemente se alquilan más servidores sin incurrir en grandes costos de infraestructura e implementación.(«*Big Data Appliance* | Oracle España», s. f.),(Cloudera, comerciales, et al., s. f.),(«AWS | Análisis de *Big Data* y almacenamiento en la nube», s. f.)

Flexibilidad: En las soluciones basadas en distribuciones tienen un amplio abanico de posibilidades para escoger el software más adecuado, al tener un ambiente menos restrictivo, bien sea en hardware y configuración como si es el caso de las soluciones tipo *appliance*, mientras que las *cloud* son básicamente de configuración.(«*Big Data Appliance* | Oracle España», s. f.),(Cloudera, comerciales, et al., s. f.),(«AWS | Análisis de Big Data y almacenamiento en la nube», s. f.)

Rendimiento: Indiscutiblemente las *appliance* son las soluciones que mejor rendimiento ofrecen ya que su hardware y software son específicos y están optimizados. Las soluciones Cloud presentan un buen rendimiento basado en las características de la infraestructura alquilada. Mientras que las distribuciones dependen de las características de los servidores adquiridos que normalmente no están optimizados.(«*Big Data Appliance* | Oracle España», s. f.),(Cloudera, comerciales, et al., s. f.),(«AWS | Análisis de *Big Data* y almacenamiento en la nube», s. f.)

Experimentación: En las soluciones de tipo *appliance* no es posible realizar pruebas antes de adquirías, en los ambientes *cloud* se da la posibilidad pero con restricciones, mientras que en las de tipo de distribuciones se pueden hacer sin inconveniente en máquinas virtualizadas.(«*Big Data Appliance* | Oracle España»,

s. f.),(Cloudera, comerciales, et al., s. f.),(«AWS | Análisis de Big Data y almacenamiento en la nube», s. f.)

Basados en los resultados expuestos en la tabla de características de tipos de soluciones *Big Data* (Ver Tabla 8) y la valoración de los parámetros anteriores, podemos decir que el tipo de solución más adecuado para iniciar un proyecto de *Big Data* en Vanguardia Liberal, desde el punto de vista económico y flexible son las soluciones de tipo distribución. Adicionalmente la infraestructura con la que cuenta el periódico actualmente soporta sin inconvenientes un montaje de estas características proporcionando un ambiente seguro para su implementación.

En un futuro si la necesidad de la empresa lo amerita nuestra recomendación es que se evalué la posibilidad de implementar un tipo de solución *Cloud*, el cual le permita crecer en capacidad de almacenamiento y procesamiento, sin la preocupación de la infraestructura de soporte. Este tipo de solución puede configurarse temporalmente si así se desea, ya que podría llegarse a necesitar solo en ciertos eventos puntuales de análisis a futuro.

5.3 Recomendaciones Gartner 2016

Como parte del proceso de comparación de las diferentes tecnologías de *Big Data*, es importante para el estudio tomar como referencia las recomendaciones de expertos sobre las mejores soluciones. En este caso vamos a tomar como base el estudio realizado por la empresa consultora y de investigación de las tecnologías de la información Gartner Group(«Acerca de Gartner», s. f.) en su estudio más reciente para el año 2016.

5.3.1 *Business Intelligence* y *Business Analytics*

El cuadrante mágico de Gartner(«Acerca de Gartner», s. f.) está compuesto por dos ejes uno horizontal y otro vertical, al hacer el análisis Ver Figura 9 se observa que el eje horizontal referencia las soluciones más completas, se aprecia que las soluciones de Microsoft aparecen en este sector, al revisar el eje vertical que analiza la habilidad de ejecución, se pueden ver empresas como *Pentaho*, que indican que sus soluciones son poco completas y tienen la desventaja de tener poca habilidad para su ejecución A pesar de ser una herramienta muy robusta en componentes que permiten disponer un gran potencial en procesos de transformación de datos, implementación de componentes *Hadoop* y escalabilidad.

De acuerdo al cuadrante mágico de Gartner las empresas con mejor tendencia durante el 2016 son: TABLEAU(«Tableau Software», s. f.), QLIK(«*Guided Analytics* | *Business Intelligence Software* | *QlikView*», s. f.) y MICROSOFT(«Power BI | Herramientas de BI para la visualización de datos interactivos», s. f.), con sus soluciones TABLEAU(«Tableau Software», s. f.), QLIKVIEW(«*Guided Analytics* | *Business Intelligence Software* | *QlikView*», s. f.) y POWER BI(«Power BI | Herramientas de BI para la visualización de datos interactivos», s. f.) respectivamente.

Empresas como IBM(«IBM - Colombia», 2016) y SAS(«SAS», s. f.) al ser empresas similares a Microsoft están ubicadas también cerca al eje vertical en la parte inferior del cuadrante indicando que también tienen deficiencias en la capacidad para ejecutar. Ver Figura 9

La organización dispone de un grupo relativamente pequeño de TI, cuyas circunstancias exigen que cualquier tipo de solución seleccionada para el caso de estudio sea lo más sencilla posible de administrar y de usar.

5.3.2 Almacenamiento y gestión de bases de datos de *Analytics*

Las empresas que representan la tendencia en almacenamiento y gestión de bases de datos para el 2016 son: ORACLE(«*Big Data Appliance* | Oracle España», s. f.), *TERADATA(«Big Data Analytics & Hadoop Services from Teradata*», s. f.), *MICROSOFT*(«Microsoft: página principal», s. f.), IBM(«IBM - Colombia», 2016) y SAP(Ferrer-Sapena & Sánchez-Pérez, 2013). Es importante ver como se configuro el cuadrante de Visionarios para esta publicación con empresas como *MapR Tecnologies*(«MapR: Plataforma de datos convergente», s. f.), Cloudera(Cloudera, comerciales, et al., s. f.), *HortonWorks*(«*Hortonworks*», s. f.), Pivotal(Software, 2015) entre otras.

Al hacer el análisis del cuadrante se puede observar que la solución de *bussines intelligence* con Microsoft será más compatible con Microsoft según este cuadrante, a pesar de que hay otras empresas mejor ubicadas como Oracle*(«Big Data Appliance* | Oracle España», s. f.) y *Teradata(«Big Data Analytics & Hadoop Services from Teradata*», s. f.), soluciones similares a las de Microsoft como AWS(«AWS | Análisis de Big Data y almacenamiento en la nube», s. f.), MongoDB(«Reinventando la gestión de datos», s. f.), Cloudera(Cloudera, comerciales, et al., s. f.) y *HortonWorks*(«*Hortonworks*», s. f.) tienen limitantes, no

son tan completas las soluciones como las de Microsoft, y no son tan fáciles de ejecutar o implementar comparándolas con las de Microsoft. Ver Figura 10.

5.4 Evaluación Distribuciones

En este apartado se realiza la evaluación de las distribuciones de *Hadoop* más extendidas e importantes del mercado actual. Esta selección se hizo basada en el cuadrante de visionarios de Gartner sobre almacenamiento y gestión de bases de datos de *analytics* 2016. Ver Figura 10.

Figura 9: Cuadrante de Gartner sobre Herramientas Business Intelligence y Business Analytics

Fuente: («Gartner BI», s. f.)

Vanguardia Liberal en sus procesos de compra de tecnología a nivel interno considera siempre la plataforma sobre la cual debe instalarse la herramienta, la

presencia en el mercado en términos de que tan usada es por otras empresas y la productividad y desarrollo en términos de que tan sencillo es operar la herramienta entre otros. Como parámetros adicionales en esta evaluación se incluyen la opción de versión gratuita, componentes *Hadoop* que incluye la herramienta, Herramientas de Administración, Componentes que mejoren el rendimiento o desempeño y la tolerancia a fallos como factores propios a evaluar en herramientas *Big Data*(Garcés Uquillas, 2015), según consultoras de renombre como Gartner(«Acerca de Gartner», s. f.), Forrester(«Forrester : *Welcome*», s. f.), *Dresner(«Dresner Advisory Services - Home of Business Intelligence and the Wisdom of Crowds ® Market Research»*, s. f.) y BARC(«*The BI Survey 15, BARC's annual report on the BI industry*», s. f.).

Figura 10: Cuadrante de Gartner sobre Almacenamiento y gestión de bases de datos de Analytics

Fuente: («Cuadrante Mágico para *Data Warehouse y Data Management Solutions* para *Analytics*», s. f.)

En el Anexo E se describe en detalle la escala utilizada para la comparación teórica de sus características y funcionalidades. Al final se realiza una

comparación que permite tener una idea de qué distribución es más conveniente dependiendo de las necesidades.

Después de haber descrito En el Anexo E cada uno de los parámetros seleccionados para realizar la evaluación comparativa entre las distribuciones, podemos apreciar la comparación de características realizada entre las distribuciones *Big Data* escogidas. Ver Tabla 9

Plataforma: La Distribución de *HortonWorks* es la única que permite la instalación de un *clúster Hadoop* en nodos con el sistema operativo *Windows Server*.(«*Hortonworks*», s. f.). Ver Tabla 9

Presencia en el mercado: Entre más presencia tenga en el mercado, mayor será la comunidad que respalda la solución y por consiguiente más información y soporte estará disponible acerca de los posibles inconvenientes que puedan surgir. Cloudera y *Hortonworks* son distribuciones muy usadas ya que la primera, aparte de ser la distribución puntera en desarrollo *Hadoop*, también es la escogida por Oracle para su solución *Big Data*. La segunda en cambio es la escogida por Microsoft y Teradata. Pivotal tiene detrás a una gran multinacional como EMC y aunque es una recién llegada, su solución no deja de ser la que ofrecía EMC antes con el nombre de *Greenplum*.(«*Hortonworks*», s. f.),(Cloudera, comerciales, et al., s. f.)

Versión gratuita: Todas las distribuciones tienen una opción gratuita, lo que las diferencia de una a otra son las restricciones para poderlas trabajar. («*Hortonworks*», s. f.),(Cloudera, comerciales, et al., s. f.),(Software, 2015),(«IBM - *InfoSphere Information Server - Data Integration, Information Integration - Overview*», s. f.),(«MapR: Plataforma de datos convergente», s. f.)

Componentes *Hadoop*: Cloudera y *HortonWorks* son las distribuciones que menos han modificado el núcleo de *Hadoop*, por lo tanto pueden adaptar de forma más rápida las nuevas versiones. Además de estas dos compañías *open source*, esta Pivotal junto a esta empresa son las únicas que ofrecen una distribución de *Hadoop* con el nuevo gestor de recursos YARN. («*Hortonworks*», s. f.),(Cloudera, comerciales, et al., s. f.)

Administración: Todas las distribuciones tienen algún tipo de herramienta de gestión y monitorización del estado del clúster, un asistente automático de instalación, control de acceso LDAP y acceso mediante API REST al estado del

clúster. («*Hortonworks*», s. f.),(Cloudera, comerciales, et al., s. f.),(Software, 2015),(«IBM - *InfoSphere Information Server - Data Integration, Information Integration - Overview*», s. f.),(«MapR: Plataforma de datos convergente», s. f.)

Tabla 9: Características de Distribuciones Big Data

Características	Cloudera	Pivotal HD	IBM *InfoSphere*	*MapR*	*Horton Works*
Plataforma	**1**	**1**	**1**	**1**	**3**
Presencia en el mercado	**3**	**2**	**2**	**3**	**3**
Versión Gratuita	**3**	**1**	**1**	**1**	**2**
Componentes *Hadoop*	**3**	**1**	**2**	**2**	**3**
Administración	**2**	**2**	**2**	**2**	**2**
Productividad y Desarrollo	**2**	**2**	**3**	**1**	**1**
Rendimiento	**3**	**3**	**2**	**2**	**2**
Tolerancia a Fallos	**2**	**2**	**2**	**3**	**2**
Total	**19**	**14**	**15**	**15**	**18**

Fuente: Autor

Productividad y desarrollo: En la categoría de productividad y desarrollo IBM ha sido la distribución que más provecho ha sabido sacar, ofreciendo un gran conjunto de herramientas que facilitan el desarrollo de aplicaciones de análisis y por lo tanto aumentan la productividad de los analistas y desarrolladores.(«IBM - *InfoSphere Information Server - Data Integration, Information Integration - Overview*», s. f.)

Rendimiento: Entre lo más destacable está la posibilidad de realizar consultas SQL interactivas de Cloudera y Pivotal HD. También es importante destacar la herramienta *Data Loader* de *Pivotal* HD que permite mover grandes volúmenes de datos en paralelo entre bases de datos ajenas al *clúster Hadoop* y el *clúster*.(Cloudera, comerciales, et al., s. f.),(Software, 2015)

Tolerancia a fallos: MapR sustituye HDFS por MapR-FS, un sistema de ficheros con una filosofía muy similar a la de Cassandra, y añadiendo alta disponibilidad en las tareas *MapReduce*.(«MapR: Plataforma de datos convergente», s. f.)

De acuerdo a los resultados obtenidos en la comparación de características de las principales distribuciones de *Big Data* (Ver Tabla 9), concluimos que la opción más favorable para implementar en la organización, es la Distribución de Cloudera. Es importante mencionar que la segunda más importante de acuerdo a este análisis corresponde a la distribución de *HortonWorks*, El punto decisorio entre las dos fue la cantidad de herramientas *open source* que ofrecía cada una en sus versiones gratuitas.

Hasta este punto hemos tomado dos decisiones muy importantes tomando como base el punto de vista económico y el flexible, dando como resultado la selección de una solución de tipo distribución como es el caso de la Distribución de Cloudera. La implementación de esta solución está proyectada para ser desarrollada en un mediano plazo que se puede visualizar en el plan de implementación, que se aprecia en el Capítulo 7

5.5 Evaluación de Herramientas de *Bussiness Intelligence*.

Como parte esencial de este trabajo de investigación está: Definir qué herramientas son las más apropiadas para ser implementadas de acuerdo a las características de la empresa y sus procesos de negocio.

Los procesos de información y las demandas actuales en cuanto a capacidad de cómputo, no exigen tener una elasticidad que haga necesario la implementación de un esquema *Hadoop* en Vanguardia Liberal. Según la literatura la mayor fortaleza de *Big Data* o de soluciones tipo *Hadoop*, es precisamente el poder atender demandas inesperadas y cambios grandísimos en requerimientos de cómputo, que actualmente no se están presentando en la organización.

Por tal motivo la recomendación para iniciar con la adopción de proyectos de *Big Data* es no abordarlo por elasticidad si no por el contrario centrarse en la necesidad apremiante, la analítica que debe incluir diferentes fuentes de información. Ahora bien para tomar la decisión más adecuada en este aspecto se tomó como referencia el cuadrante de Gartner sobre herramientas *Business Intelligence* y *Business Analytics* del año 2016. Ver Figura 9

Tomando las recomendaciones dadas por Gartner en su Cuadrante Mágico, se decidido realizar un análisis comparativo basado en parámetros de evaluación que Vanguardia Liberal considera siempre en sus procesos de compra de tecnología como: Costo de licenciamiento, Curva de aprendizaje, Tiempo de implementación,

Socios de negocio entre otros. Como parámetros adicionales en esta evaluación se incluyen la comunidad de respaldo, procesos de transformación de datos, conectividad a fuentes de datos e integración con componentes *Hadoop* como factores propios a evaluar en herramientas *Big Data*(Garcés Uquillas, 2015), según consultoras de renombre como Gartner(«Acerca de Gartner», s. f.), Forrester(«Forrester : Welcome», s. f.), Dresner*(«Dresner Advisory Services - Home of Business Intelligence and the Wisdom of Crowds ® Market Research*», s. f.) y BARC(«*The BI Survey 15, BARC's annual report on the BI industry*», s. f.).

Las empresas que conforman este análisis fueron seleccionadas básicamente por su ubicación en el cuadrante de líderes. Adicionalmente se toman dos de las empresas más representativas de *Bussiness Intelligence* relacionadas con código abierto, que se ubican en el cuadrante de visionarios. En el Anexo F se describe en detalle la escala de valores utilizada para ponderar los parámetros seleccionados. A continuación podemos apreciar la tabla comparativa de características de las herramientas de inteligencia de negocios seleccionadas. Ver Tabla 10

Tabla 10: Características de Herramientas de Inteligencia de Negocios

Características	**Tableau**	***Qlik (QlikView)***	***Microsoft (Power BI)***	***Pentaho BI***	***Tibco (SpotFire)***
Costo licencia *Cloud* x Mes	**1**	**1**	**2**	**3**	**1**
Comunidad de Respaldo	**2**	**3**	**3**	**2**	**2**
Curva de Aprendizaje	**3**	**2**	**2**	**1**	**2**
Tiempos de Implementación	**3**	**2**	**3**	**2**	**2**
Socios de Negocio	**2**	**2**	**3**	**2**	**3**
Transformación de Datos	**2**	**1**	**2**	**3**	**2**
Conectores a Fuentes de Datos	**2**	**2**	**1**	**3**	**2**
Componentes *Hadoop*	**2**	**2**	**2**	**3**	**2**
Total	**17**	**15**	**18**	**19**	**16**

Fuente: Autor

Costo licencia: A nivel de costos de licenciamiento se tomó como referencia los costos de una licencia tipo *cloud* por mes. En este aspecto se puede decir que la mejor opción es *Pentaho BI*, ya que dispone de una distribución gratuita. Como segunda opción estaría la herramienta de Microsoft. La cual tiene unos costos de 9.99 USD por licencia. Siendo la herramienta propietaria más económica en el mercado actualmente.(«Power BI | Herramientas de BI para la visualización de datos interactivos», s. f.)

Comunidad de respaldo: La comunidad de usuarios de cada una de las herramientas es algo que influye notablemente en la elección de cualquier tecnología actualmente ya que a través de ellas se llegan a resolver múltiples casos de una forma sencilla y económica. Las herramientas que mejor ofrecen respaldo debido a su gran cantidad de usuarios que utilizan la herramienta son *Power BI* y *QlickView* con más de 500 mil usuarios.(«*Guided Analytics* | *Business Intelligence Software* | *QlikView*», s. f.),(«Más de 500 mil usuarios únicos de 45 mil empresas en 185 países ayudaron a darle forma al nuevo *Power BI* | *News Center* Latinoamérica», s. f.)

Curva de Aprendizaje: Según el último informe de Gartner publicado en febrero de 2016.(«Gartner BI», s. f.) menciona que la facilidad con la que se aprende a utilizar una herramienta de inteligencia de negocios sea convertido en un factor determinante para inclinarse por una u otra herramienta. En este aspecto las que menos trabajo cuesta de aprender es *Tableau*.(*SelectHub*, 2016)

Tiempos de Implementación: Para tener un punto de comparación más unificado en este aspecto se analizó la implementación de la herramienta en ambientes *cloud*, lo cual reduce los tiempos de implementación considerablemente. Bajo estas condiciones las herramientas que representan tiempos relevantes son *Tableau* y *Power BI*, con unos tiempos que oscilan en menos de una semana para ver resultados.(*SelectHub*, 2016)

Socios de Negocio: Otro factor muy importante es tomar la decisión correcta al escoger la empresa que será el intermediario entre el proveedor y el cliente. Entre más opciones existan para tomar esta decisión más respaldo se puede tener ante cualquier requerimiento. Los productos con más representantes autorizados en Colombia son *Power BI* y *SpotFire*.(«*Tableau Software*», s. f.),*(«Guided Analytics* | *Business Intelligence Software* | *QlikView*», s. f.),(«*Power BI* | Herramientas de BI para la visualización de datos interactivos», s. f.),(«*Pentaho* | *Data Integration and*

Business Analytics Platform for Big Data Deployments», s. f.),(«*Data Visualization & Analytics Software - TIBCO Spotfire*», s. f.)

Transformación de Datos: Dadas las necesidades de integración de diferentes fuentes de datos a ser analizadas los procesos ETL que pueden brindar las herramientas se convierten en un factor muy importante. En este aspecto la suite de *Pentaho* es la que mayores fortalezas presenta en este sentido gracias al componente *Kettle*.(«*Pentaho* | *Data Integration and Business Analytics Platform for Big Data Deployments*», s. f.)

Conectores a fuentes de datos: Una característica fundamental en este proyecto particular es la necesidad de conectividad a múltiples fuentes de datos que permitan la integración de las diferentes fuentes de información del periódico. *Kettle* es un componente incluido en la suite de *Pentaho* que permite de una forma muy fácil la conexión a diferentes orígenes de datos.(«*Pentaho* | *Data Integration and Business Analytics Platform for Big Data Deployments*», s. f.)

Componentes Hadoop: Proyectando la necesidad de implementar más funcionalidades que ofrecen las múltiples herramientas que conforman el ecosistema de *Hadoop* es muy importante la facilidad de integración de componentes de este tipo. En este aspecto *Pentaho* por ser una herramienta de código abierto tiene una ventaja importante sobre las demás.(«*Pentaho* | *Data Integration and Business Analytics Platform for Big Data Deployments*», s. f.)

Según el análisis realizado sobre las herramientas propuestas por Garthner como líderes y visionarias en el cuadrante mágico publicado en febrero de 2016, acerca de las herramientas de inteligencia de negocios.(«Gartner BI», s. f.) Podemos decir que existe un empate técnico entre *Power BI* de Microsoft y *Pentaho BI*. Fácilmente la decisión podría inclinarse por Microsoft si tenemos en cuenta que actualmente Vanguardia Liberal cuenta con el licenciamiento de otras herramientas Microsoft como es el motor de bases de datos SQL *Server Enterprise* 2008 R2. Siendo esta herramienta su complemento natural.

Sin embargo teniendo en cuenta las condiciones de integración, procesamiento, limpieza y transformación de datos que la empresa va a requerir a futuro, se sugiere como la opción más indicada la suite de *Pentaho BI*, dadas sus características de elasticidad, escalabilidad, integración, transformación y múltiples conectores a datos que ofrece además de su integración con herramientas del ecosistema de *Hadoop*. Tambien es bien importante mencionar que esta

herramienta brinda componentes de visualización y de minería de datos muy importantes y pertinentes entre los aspectos analizados en la encuesta desarrollada a Gerentes, Sub Gerentes y Coordinadores de las unidades de Publicidad, Circulación, Administración y Tecnología del periódico.

5.6 Consideraciones Técnicas.

Después de haber revisado y evaluado varias tecnologías de *Big Data*, es el momento de cuestionarnos con relación a si es apropiado actualmente implementar un ambiente de *Hadoop*, teniendo en cuenta que la principal razón que lleva a las organizaciones a implementar un esquema de este tipo tiene que ver con la elasticidad del negocio y la demanda de análisis que se exige en el acontecer diario de la empresa, con puntos altos en donde se debe atender la demanda adicional y momentos planos en los cuales no es necesario contar con tantos recursos, en ese aspecto Vanguardia Liberal actualmente no tiene demandas de computo ni de almacenamiento que determinen la implementación de un ambiente de elástico.

Todas las empresas quieren usar *Big Data*, todos quieren estar en ese escenario y todos quieren sentar el precedente que utilizan este tipo de tecnologías. Conscientemente se revisaron las necesidades del negocio y se aprecia en la literatura que la mayor fortaleza de *Big Data* o de soluciones tipo *Hadoop*, tiene que ver con que se atienden las demandas inesperadas y cambios grandísimos en requerimientos de computo, las cuales Vanguardia Liberal no experimenta actualmente.

Por otra parte se evidencia la necesidad de avanzar en el proceso de maduración de análisis y visualización de datos, de ahí la importancia de adoptar una tecnología que reúna las mayores condiciones necesarias que brinden el avance en estos aspectos. Por esta razón incluimos directamente el análisis de una herramienta de Inteligencia de Negocios. De igual forma la propuesta realizada para la adopción de una solución de *Big Data* como es el caso de la distribución de Cloudera, sigue siendo totalmente valida de implementar a un mediano plazo, volviendo a evaluar en su debido momento las exigencias de elasticidad que tenga la empresa, también será pertinente evaluar en ese momento la viabilidad de migrar a una infraestructura *Cloud*. En la Tabla 11 se puede apreciar un resumen de las tecnologías propuestas, características principales por las cuales se consideraron como las más apropiadas, objetivo principal que su implementación llegaría a resolver para este proyecto en particular.

Tabla 11: Resumen Tecnologías Big Data propuestas

Tecnologías Big Data Propuestas		
Tecnología	**Características**	**Objetivo**
Distribución de Cloudera	• *Open Source* • Mayor trayectoria en ambientes *Hadoop* • Implementa la mayor cantidad de componentes *Hadoop* • Es un paquete integrado que contiene múltiples herramientas que facilitan la gestión • Satisface las necesidades de integración, Carga, transformación de datos de diferentes fuentes • Brinda la posibilidad de múltiples conectores a fuentes de datos • Dispone de procesos de analítica y estadística	• Satisfacer las necesidades de un ambiente elástico evaluable en un mediano plazo • Resolver las necesidades de analítica y visualización • Unificar la información mediante los procesos de transformación de datos
Pentaho BI	• *Open source* • Dispone de componentes especializados en el cargue de datos • Múltiples conexiones a fuentes de datos • Procesos de transformación de datos • Incluye herramientas de analítica y visualización de datos	• Dar solución a las necesidades de analítica y visualización en un corto plazo • Integrar la información de las diferentes fuentes de datos • Permitir el análisis de la información unificada • Presentar los diferentes cuadros de mando

Fuente: Autor

6. ARQUITECTURA Y MODELOS PROPUESTOS

En este capítulo se revisan tres modelos de arquitectura de *Big Data* y se plantea uno basado en las características propias del negocio tomando los elementos más adecuados de las tres revisadas, proponiendo así una arquitectura nueva que pueda ser referente para implementaciones de proyectos de *Big Data* en otras empresas sin importar el sector económico al que pertenezcan. Posteriormente basados en la arquitectura propuesta se presenta un modelo que implemente las determinaciones tomadas en el capítulo 3, buscando mejorar la formulación de estrategias comerciales y de segmentación en Vanguardia Liberal.

Se revisaron varias arquitecturas de referencia entre las cuales estaban la de IBM(Rodriguez & Valverde, s. f.), Oracle(Rodriguez & Valverde, s. f.), Cloudera(Rodriguez & Valverde, s. f.), Microsoft(Vega, Ortega, & Aguilar, 2015), Krishnan Kris(Vega et al., 2015) y Bob Marcus(Vega et al., 2015).

Teniendo en cuenta las necesidades del periódico la arquitectura a implementar debe ser abierta y sencilla de tal forma que permita incluir cualquier componente del ecosistema *Hadoop*. Con este criterio claro las arquitecturas de referencia de IBM(Rodriguez & Valverde, s. f.), Oracle(Rodriguez & Valverde, s. f.) y Cloudera(Rodriguez & Valverde, s. f.), fueron descartadas ya que involucraban la implementación de las mismas con herramientas propietarias o definían un esquema muy particular, la de Microsoft no se descartó por ser la opción más natural a realizar, ya que se dispone del licenciamiento del motor de la base de datos SQL *server* 2008 R2, junto a las propuestas de Krishnan Kris*(«Data Warehousing in the Age of Big Data, 1st Edition* | Krish Krishnan | ISBN 9780124059207», s. f.) y Bob Marcus(Vega et al., 2015) se procedió a su revisión y análisis.

6.1 Arquitectura de procesamiento de *Big Data* propuesta por Krishnan Krish.

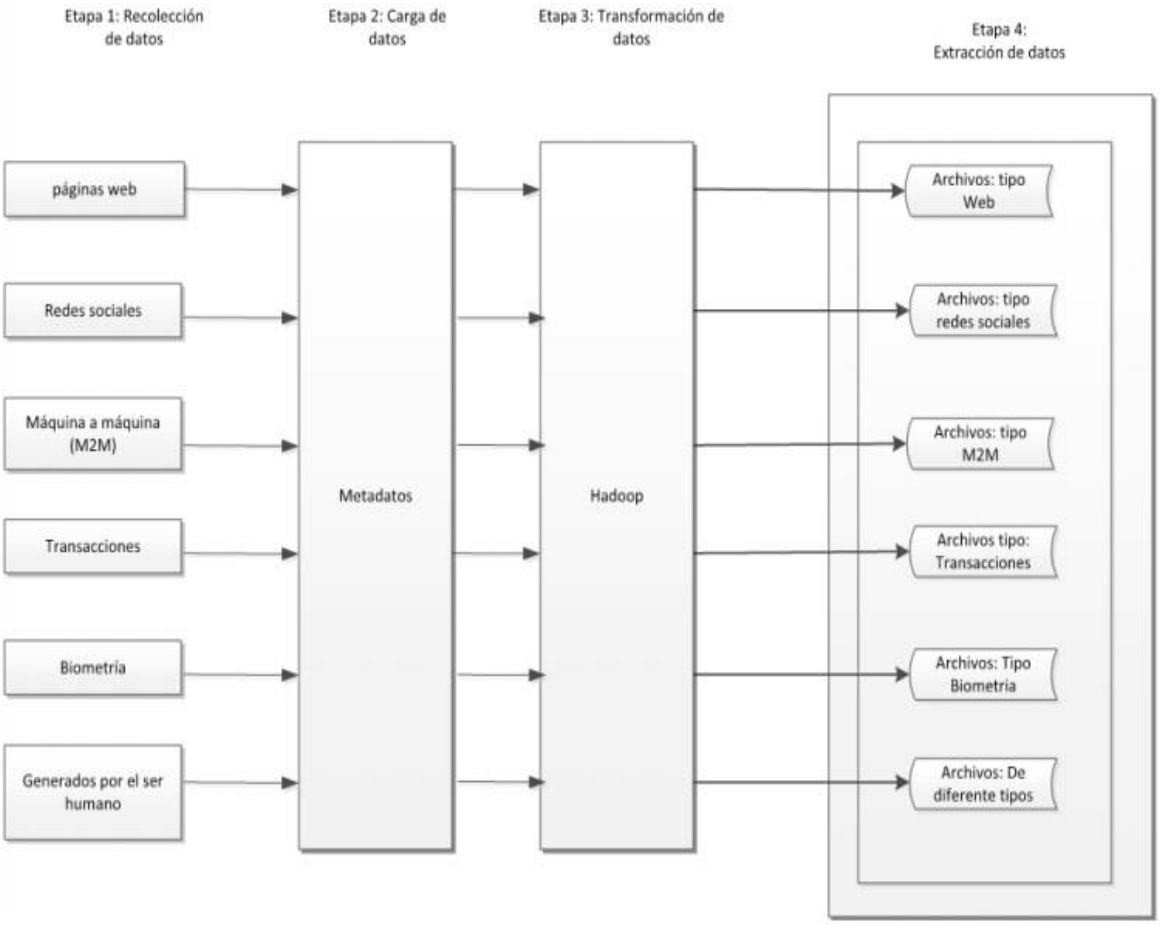
La arquitectura planteada por el Krish define cuatro (4) etapas: Recolección o recopilación, carga, transformación y extracción de datos. Ver Figura 11

Etapa 1. Recolección de datos: En la primera etapa, los datos son recibidos de diferentes orígenes o fuentes, que pueden ser: páginas web, redes sociales,

máquina a máquina (M2M), transacciones, biometría o generados por el ser humano.(Vega et al., 2015)

Etapa 2. Carga de datos: En esta etapa, los datos se cargan aplicando el concepto de metadatos (datos que describen otros datos). Además de la carga como tal, es la primera vez que los datos se estructuran.(Vega et al., 2015)

Figura 11: Arquitectura de Big Data propuesta por Krish



Fuente: (Vega et al., 2015)

Etapa 3. Transformación de datos: En este punto, los datos se transforman mediante la aplicación de las reglas del negocio y el procesamiento de los datos. Respecto al procesamiento, en cada etapa producen resultados intermedios que se pueden almacenar para un posterior examen. El resultado de esta etapa son unas cuantas claves de metadatos con modelo clave-valor.(Vega et al., 2015)

Etapa 4. Extracción de datos: El objetivo de la extracción es obtener datos para su posterior análisis, generar informes operativos y su posible visualización y almacenamiento.(Vega et al., 2015)

6.2 Arquitectura de *Big Data* propuesta por Bob Marcus.

La arquitectura planteada por el autor define siete (7) niveles: Fuentes de datos externos, Secuencia y procesamiento ETL, Fundación altamente escalable, Bases de datos operacionales y de analítica, Análisis e interfaces de bases de datos, Aplicaciones e interfaces de usuario. Ver Tabla 12.

Tabla 12: Arquitectura Big Data propuesta por Marcus

F. Aplicaciones e interfaces de usuario	**G. Servicios de Apoyo**
E. Análisis e interfaces de bases de datos	
D. Bases de datos operacionales y de analítica	
C. Fundación altamente escalable	
B. Secuencia y procesamiento ETL	
A. Fuentes de datos externos	

Fuente: (Vega et al., 2015)

Nivel A. Fuentes de datos externas: Es parte de la arquitectura de datos que suministra las entradas de datos externas y la producción de los componentes internos de *Big Data*.(Vega et al., 2015)

Nivel B. Secuencia y procesamiento ETL: Filtra y transforma los flujos de datos provenientes de los recursos externos.(Vega et al., 2015)

Nivel C. Fundación altamente escalable: Existen tres tipos de escalamiento:

- A nivel de la infraestructura, existe con el fin de poder atender el almacenamiento y procesamiento de grandes volúmenes de datos.(Vega et al., 2015)

- Referente a los almacenes de datos. Tal como lo menciona Marcus, “es la esencia de la arquitectura *Big Data*”, la cual sucede en forma de “escalabilidad horizontal que usando componentes menos caros puede apoyar el crecimiento ilimitado de almacenamiento de datos”.(Vega et al., 2015)

- Escalonar el procesamiento mediante el procesamiento distribuido en paralelo escalable con tolerancia a fallos similares.(Vega et al., 2015)

Nivel D. Bases de datos operacionales y de Analíticas: Propone tres clases de bases de datos:

- Bases de datos analíticas. El análisis de bases de datos toma los datos procesados y escalonados de la sección anterior. Son bases de datos altamente optimizadas para sola lectura (por ejemplo, columnas de almacenamiento, amplia indexación y des normalización). A menudo es aceptable para las respuestas de base de datos por tener una latencia alta (por ejemplo, invocar el procesamiento por lotes escalable sobre grandes conjuntos de datos). (Vega et al., 2015)

- Bases de datos operacionales. Estas bases de datos mantienen una excelente operación en lectura y escritura en general de forma eficiente. Por ejemplo, las bases de datos NoSQL, son de uso frecuente en las arquitecturas de datos grandes en esta capacidad. Los datos pueden ser posteriormente transformados y cargados en las bases de datos analíticas para soportar aplicaciones analíticas.(Vega et al., 2015)

- En la memoria de datos Grids. Se refieren a los datos ubicados en memorias cachés, que buscan minimizar escribir en disco los datos. Se pueden utilizar para aplicaciones en tiempo real a gran escala que requieren acceso transparente a los datos.(Vega et al., 2015)

Nivel E. Analítica e interfaces de bases de datos: Está compuesto por tres partes:

- Análisis de interfaces de procesos en lotes. Se refiere al tipo de interfaz usada para el procesamiento de datos que provienen en lotes o Batch (p. ej. *Map-Reduce*). También hace referencia a la interfaz de usuario para acceder a los datos en almacenes de datos escalables (por ejemplo del sistema de archivos *Hadoop*).(Vega et al., 2015)

- Análisis de interfaces interactivas. “Los almacenes de datos pueden ser bases de datos escalables horizontalmente sintonizados para las respuestas interactivas (p. ej. *HBase*) o lenguajes de consulta en sintonía con los modelos de datos”.(Vega et al., 2015)

- Análisis de interfaces en tiempo real. Se deben analizar con cuidado las interfaces que atienden o son parte de un sistema de tiempo real, pues los eventos son complejos tanto en el procesamiento como almacenamiento de los datos.(Vega et al., 2015)

Nivel F. Aplicaciones e interfaz de usuario: Se refiere a las aplicaciones e interfaces de usuario, las cuales no deben ser algoritmos complejos, al usar grandes cantidades de datos distribuidos.(Vega et al., 2015)

Nivel G. Servicios de apoyo: Tienen que ver con los componentes necesarios para la implementación y gestión de sistemas robustos de *Big Data*, los cuales se pueden discriminar como:

- Diseñar, desarrollar e implementar herramientas de alto nivel de calidad, de manera que sirvan para implementar soluciones *Big Data*.(Vega et al., 2015)

- Seguridad. Aspecto importante a nivel de controles de seguridad de grandes volúmenes de datos, pues en la actualidad son escasos o limitados.(Vega et al., 2015)

- Gestión de procesos. “Los distribuidores comerciales son el suministro de herramientas de gestión de procesos para aumentar las implementaciones de código abierto”.(Vega et al., 2015)

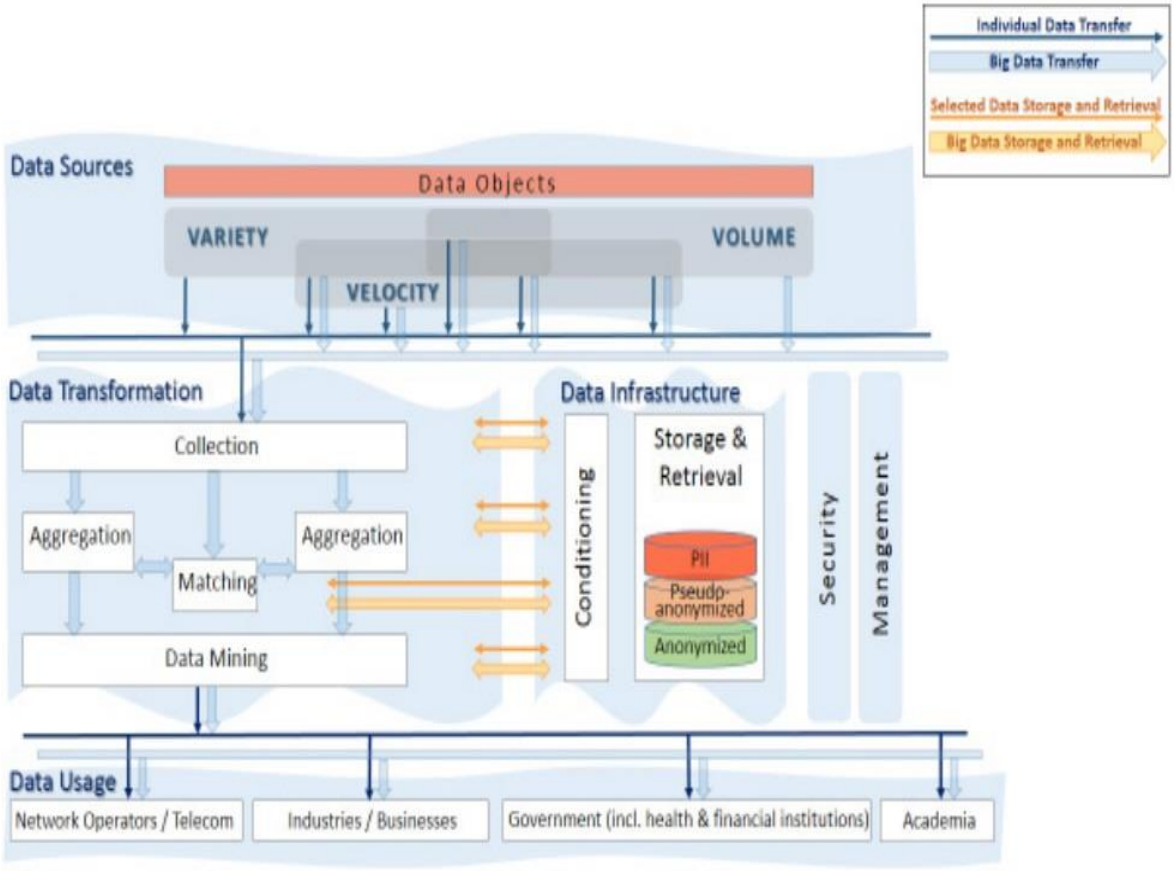
- Gestión de recursos de datos. Tiene relación con las “Herramientas de control de datos de código abierto que son todavía inmaduras. Estos serán aumentados en un futuro por los proveedores comerciales”.(Vega et al., 2015)

6.3 Arquitectura de *Big Data* propuesta por Microsoft.

Está compuesta por cuatro componentes: fuentes de datos (*Data Sources*), transformación de datos (*Data Transformation*), infraestructura de datos (*Data Infrastructure*) y uso de datos (*Data Usage*).(Vega et al., 2015) Ver Figura 12.

Componente 1. Fuentes de datos: Estos datos presentan tres características que definen el *Big Data*: volumen, velocidad y variedad. Como aspecto relevante, este tipo de datos son independientes de su contenido o del contexto. “Por lo general, los datos de ‘*Big Data*’ se recogen para un propósito específico”. Además, “una vez que se recogen los datos, se pueden volver a utilizar para una variedad de propósitos, algunos potencialmente desconocidos en el tiempo de recogida”.(Vega et al., 2015)

Figura 12: Arquitectura Big Data propuesta por Microsoft



Fuente: (Vega et al., 2015)

Componente 2. Transformación de datos: Consta de cuatro sub etapas. Cada una de ellas puede tener su etapa de pre procesamiento específico, creación de metadatos, utilizar diferentes infraestructuras de datos especializados de acuerdo con sus necesidades, tener su propia privacidad y, por último, sus propias políticas. (Vega et al., 2015)

- **Colección:** En la transformación de datos aparece un proceso llamado colección, donde se recogen datos en diferentes tipos y formas, se pueden allegar datos de fuentes y estructuras similares o iguales, o combinadas. Además, se crean metadatos para que sea más fácil hacer una búsqueda de los datos. (Vega et al., 2015)

- **Agregación:** Consiste en adicionar datos a una colección más grande, cuando uno o varios metadatos tengan claves iguales. “Como resultado, la información acerca de cada objeto se enriquece o el número de objetos en la colección crece”.(Vega et al., 2015)

- **Congruencia:** En esta etapa se recogen datos con metadatos sin importar si son diferentes y se unen a una colección más grande. Al final se obtiene que cada objeto sea enriquecido.(Vega et al., 2015)

- **Minería de datos:** La minería o *data mining* es una extracción de datos para luego poder hallar relaciones entre ellos. Existen “dos tipos de minería de datos: descriptivo, que proporcione información sobre los datos existentes, y predictivo, lo que hace que los pronósticos basados en los datos”.(Vega et al., 2015)

Componente 3. Infraestructura de *Big Data*: Se considera la infraestructura como “un paquete de almacenamiento de datos o software de base de datos, servidores, almacenamiento y redes utilizados en apoyo de las funciones de transformación de datos y de almacenamiento de datos según sea necesario”.(Vega et al., 2015)

Componente 4. Uso de los datos: Depende del usuario y sus necesidades particulares, pero estos se pueden presentar en diferentes formatos y bajo ciertas consideraciones de seguridad.(Vega et al., 2015)

6.4 Comparación de Arquitecturas Krishnan, Marcus y Microsoft.

Al comparar las tres arquitecturas en forma simultánea (Ver Tabla 13) se puede evidenciar una desigualdad entre las tipificaciones (Etapas, Niveles, Componentes) que cada autor le da a la agrupación de procesos o actividades que sugiere. Sin embargo al revisar más detenidamente, se aprecia generalidad en las diferentes capas definidas de algunas arquitecturas y en otras un mayor detalle. Es decir que cada una de las propuestas plantea en términos generales un

mismo enfoque salvo ciertos aspectos específicos de cada una de las arquitecturas mencionadas.

En el modelo de Marcus aparece una etapa “(G) Servicios de apoyo”, de la cual es importante destacar el aspecto “seguridad”, que no es contemplado en las otras arquitecturas. Diseñar, desarrollar e implementar herramientas, no debe ser parte de la labor de la persona que implementa un proyecto de *Big Data*, pues hoy existen herramientas que ayudan con esta gestión. Es decir, no es necesario seguir en este aspecto el modelo de Marcus. Finalmente, la “Gestión de recursos de datos”, donde menciona las “Herramientas de control de datos de código abierto que son todavía inmaduras”; no se debe tener presente para abordar proyectos de *Big Data*, pues realmente lo que se necesita son herramientas para *Big Data* que muestren altos niveles de calidad. Por estas razones excluimos este nivel de la arquitectura propuesta por Marcus del cuadro comparativo de arquitecturas de *Big Data*. Ver Tabla 13.

Tabla 13: Cuadro Comparativo de Arquitecturas Big Data

<table>
<tr><th>Krishnan</th><th>Microsoft</th><th>Marcus</th></tr>
<tr><td>Recolección de datos</td><td>Fuentes de datos</td><td>Fuentes de datos externas</td></tr>
<tr><td>Carga de datos</td><td rowspan="2">Transformación de datos</td><td rowspan="2">Secuencia y procesamiento ETL</td></tr>
<tr><td>Transformación de datos</td></tr>
<tr><td rowspan="4">Extracción de datos</td><td rowspan="2">Infraestructura de Big Data</td><td>Fundación altamente escalables</td></tr>
<tr><td>Bases de datos operacionales y analíticas</td></tr>
<tr><td rowspan="2">Uso de los datos</td><td>Analítica e interfaces de bases de datos</td></tr>
<tr><td>Aplicaciones e interfaz de usuario</td></tr>
</table>

Fuente: Autor

6.5 Arquitectura Propuesta

Después de revisar las tres propuestas de arquitectura de *Big Data*, se tomó lo mejor de cada modelo para presentar una propuesta de arquitectura, con el fin de buscar una futura implementación con el menor esfuerzo por parte de las personas encargadas de dicha labor. En la Tabla 14 Se refleja el planteamiento realizado para la implementación de una arquitectura de *Big Data* en Vanguardia Liberal.

Tabla 14: Arquitectura Propuesta Big Data

Etapa	Descripción	Arquitectura
Recolección de datos	Describe de donde provienen los datos que van a alimentar la Arquitectura, estos pueden ser de distintas fuentes y tipos de datos: 1. Estructurados: • RDBMS (*SQL Server, Sysbase, DB2, PosgreSQL, MySql*) • *DataWarehouses* • CRM • ERP 2. Semi-estructurados: • Datos generados por maquinas • XML • JSON • EDI • Correo electrónico 3. No Estructurados: • Redes sociales • Multimedia • Texto • Datos análogos • GPS	Krishnan, Marcus

Etapa	Descripción	Arquitectura
Carga de Datos	Propone el Cargue de los datos y su estructuración, es decir hace uso de *Hadoop*.	Krishnan
Transformación de Datos	Aplica cuatro sub tareas que son: • Colección: Recoge datos en diferentes formas y tipos • Agregación: Permite agregar datos a la colección ya existente • Congruencia: Permite unir datos con metadatos • Minería de Datos: Después de extraer datos, permite hallar relaciones entre los datos.	Microsoft
Extracción de Datos	Después de extraer los datos, se tienen listos para su análisis, también para generar informes operativos, su posible visualización y, por último, el almacenamiento.	Krishnan

Fuente: Autor

6.6 Diagramas de Arquitecturas Propuestas

Ahora es importante entrar a retomar los resultados de la encuesta realizada a los Sub Gerentes y Coordinadores de las unidades de negocio de Circulación, Publicidad, Administración y Tecnología. Estos resultados indican una clara demanda de implementar procesos de analítica y visualización. Por otra parte están los resultados que arroja la investigación acerca de las mejores opciones para implementar un proyecto de *Big Data*. Se plantean entonces dos escenarios: El deseado y el realizable, descritos a continuación.

6.6.1 Diagrama de Arquitectura Deseado

En la arquitectura planteada en el diagrama (Ver Figura 13) se hace una propuesta de implementación de herramientas incluidas en la distribución de Cloudera, de acuerdo a la decisión tomada en el capítulo 5. Esta arquitectura está planteada para un escenario que incluya un esquema 100% *Hadoop* que exija una plataforma elástica a nivel de capacidades de computo representadas en almacenamiento y procesamiento.

Actualmente las necesidades de Vanguardia Liberal no son elásticas en términos de capacidades de cómputo. Además este escenario tiene limitaciones relacionadas con los costos de implementación, soporte en Colombia y madurez del equipo de tecnología. Basados en lo anterior es necesario contemplar un escenario realizable, descrito en el siguiente apartado.

6.6.2 Diagrama de Arquitectura Realizable

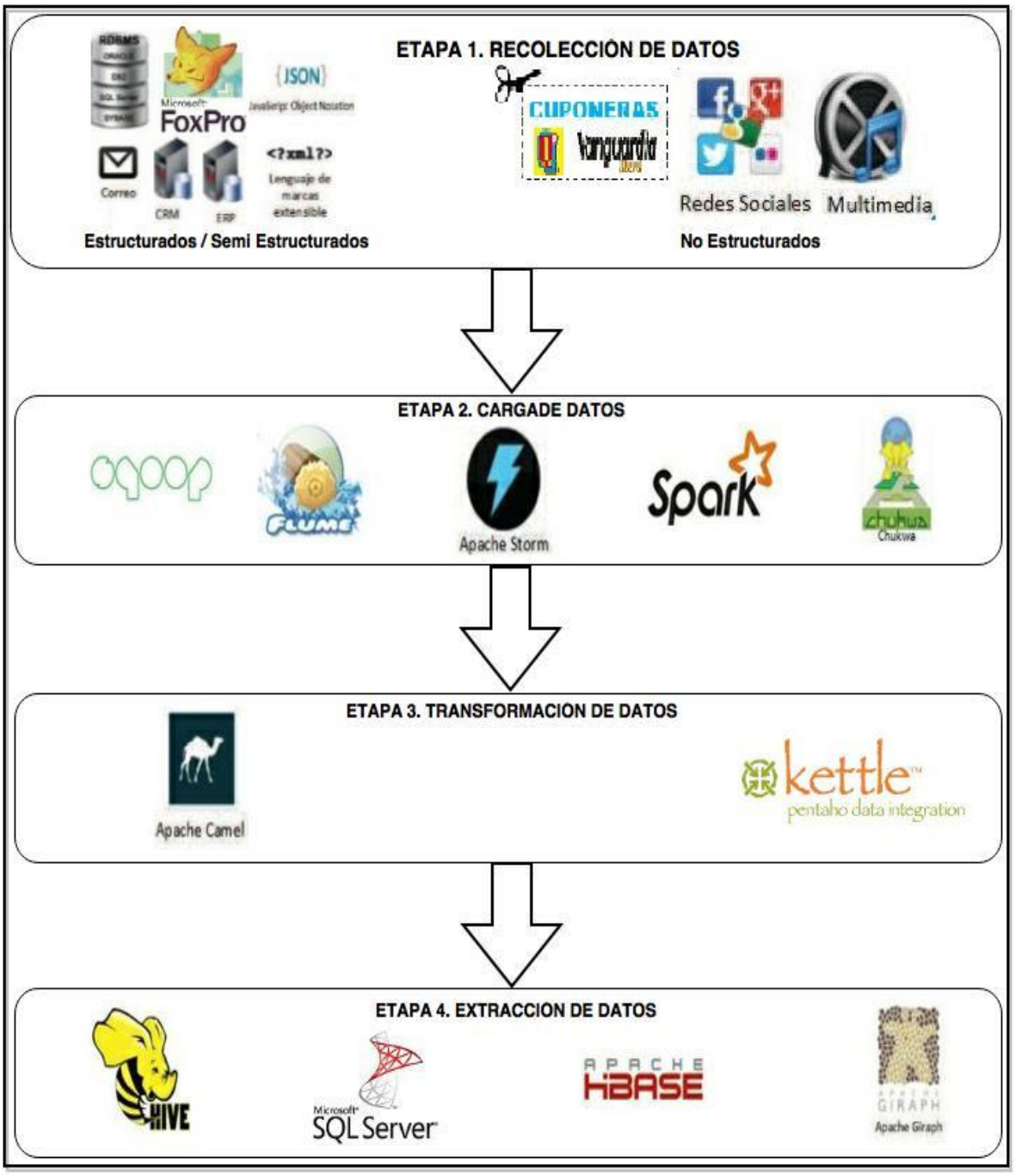
Basados en el nivel de madurez de la empresa para abordar proyectos de analítica, las demandas actuales de capacidad de computo, infraestructura y teniendo presente los resultados arrojados por la encuesta, que indican la necesidad de implementar procesos de análisis de información y visualización. Se plantea la siguiente arquitectura abordando una tecnología que cubra las necesidades requeridas por la empresa. Ver Figura 14

Este es el escenario considerado como realizable dadas las necesidades actuales de Vanguardia Liberal, se ha seleccionado la herramienta de *Pentaho Comunnity Edition* como la mejor opción para cubrir las demandas de analítica y visualización, donde es posible seguir manteniendo el esquema de almacenamiento que hay hoy en la empresa. En este escenario se renuncia a la elasticidad propia de un ambiente *Hadoop*, dadas las circunstancias actuales antes descritas en el apartado 6.6.1

La implementación de esta arquitectura dará solución al inconveniente que adolece actualmente la organización relacionado con la información parcializada que se tiene desde cada uno de las unidades de negocio con relación al cliente. Este planteamiento permite acceder a la vista 360 del cliente, ya que el proceso que pretende iniciar esta arquitectura tiene que ver con la integración de la información de los sistemas de publicidad y circulación. Por consiguiente los procesos de estrategias comerciales y segmentación se mejorarán con esta integración de la información. La proyección realizada para disponer del

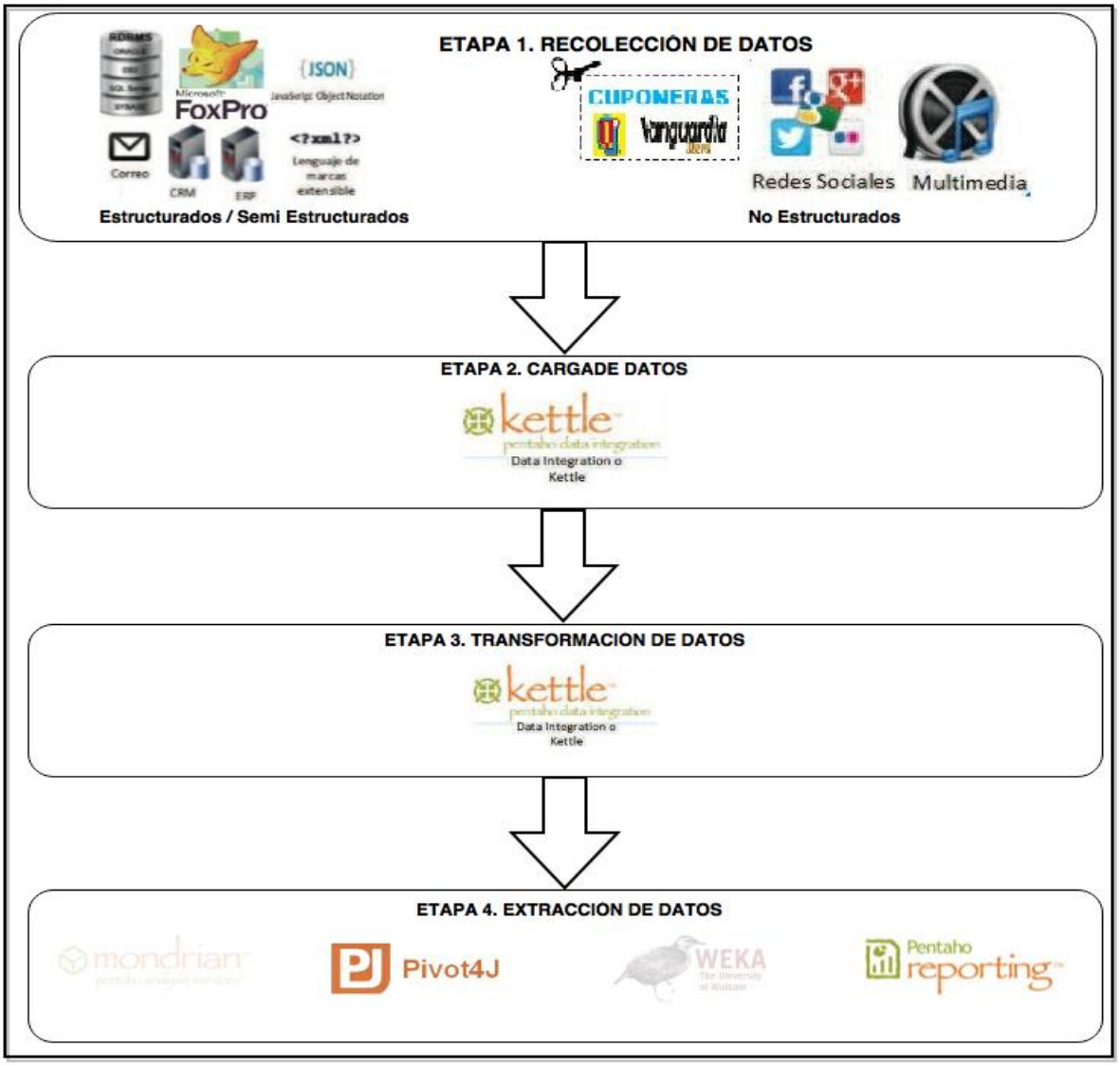
planteamiento realizado está a un corto plazo, en el capítulo 7 se puede revisar la propuesta para llevar a cabo la implementación de esta arquitectura.

Figura 13: Diagrama de Arquitectura Big Data Deseado



Fuente: Autor

Figura 14: Diagrama de Arquitectura Big Data Realizable.



Fuente: Autor

Las herramientas planteadas en el capítulo 5, para cubrir las necesidades de analítica y visualización fueron *Power BI* y *Pentaho BI*. La decisión final se tomó basada en la robustez y necesidades adicionales que se anticipan, van a existir. Esto nos indica que la mejor decisión es *Pentaho BI*. Esta herramienta es una suite, que como se puede apreciar encaja perfectamente en la arquitectura planteada. *Pentaho BI* cuenta con dos tipos de licenciamiento, una versión *Community Edition* (CE) bajo la licencia GPL y libremente descargable y la versión

Enterprise Edition (EE), basada en un modelo de suscripción, que ofrece soporte, servicios y mejoras de productos a través de la suscripción anual.(«*Pentaho* | *Data Integration and Business Analytics Platform for Big Data Deployments*», s. f.)

Pentaho es una plataforma que puede conectarse con múltiples fuentes de datos. En la etapa 2 y 3 de la arquitectura la herramienta que cubre las necesidades descritas es *Kettle*, que es una herramienta poderosa de cargue y transformación de datos que incluye la suite de *Pentaho Community*.*(«Pentaho* | *Data Integration and Business Analytics Platform for Big Data Deployments*», s. f.)

La etapa 4 es soportada por más de una herramienta incluida en la *suite*: *Mondrian* es un servidor que permite realizar procesamiento analítico en línea (OLAP), *JPivot* es una librería de componentes JSP que se utiliza para construir tablas OLAP generadas de forma dinámica.(«*Pentaho* | *Data Integration and Business Analytics Platform for Big Data Deployments*», s. f.) *Weka* es un conjunto completo de herramientas para aprendizaje automático y minería de datos. Incluye herramientas para realizar transformaciones sobre los datos, tareas de clasificación, regresión, reglas de asociación y de algoritmos de agrupamiento, se puede utilizar para ayudar a entender mejor el negocio y también mejorar el rendimiento futuro a través de análisis predictivo.(«*Weka 3 - Data Mining with Open Source Machine Learning Software in Java*», s. f.) Finalmente tenemos la herramienta *Reporting*, la cual es un conjunto de herramientas de informes *open source* que permiten crear informes relacionales y de análisis de una amplia gama de fuentes de datos y tipos de salida.*(«Pentaho* | *Data Integration and Business Analytics Platform for Big Data Deployments*», s. f.)

Los componentes de la suite que entrarían a ser piezas fundamentales del proceso son *Kettle, Pivot4J* y *Reporting*, con los cuales daríamos solución a las necesidades planteadas, los demás componentes expuestos serian evaluados en etapas posteriores.

6.7 Modelos Propuestos

En este apartado se plantean los modelos de análisis, de alto nivel que deben ser luego validados e implementados por Vanguardia liberal, para mejorar sus procesos de estrategia comercial. Se plantean básicamente dos procesos el primero corresponde al modelo integrador de la información y el segundo hace referencia al modelo de ventas unificado que deberá adoptar a corto plazo la empresa.

6.7.1 Modelo Integrador de Información

Este modelo busca la unificación de la información más importante para cualquier empresa, la del cliente, en una vista consolidada y visible por toda la fuerza de ventas de la empresa sin importar la unidad de negocio a la que pertenezca. Este modelo de integración busca acabar con el fenómeno de las “islas de información”(Coronel, 2011) que actualmente padece la organización. A continuación se puede apreciar el modelo. Ver Figura 15.

Este modelo plantea tres objetivos principales el primero tiene que ver con la centralización de los clientes de Vanguardia Liberal, logrando con esto la unificación y eliminando el fenómeno de las islas de información. El segundo objetivo busca que los sistemas legados sean dependientes, forzando a tener un solo punto de ingreso y actualización de los datos del cliente logrando que la información del cliente sea única en cualquiera de los sistemas de la empresa. Y por último lograr obtener la vista 360 del cliente que permitirá a la fuerza de ventas y a la empresa tener siempre una visión holística del comportamiento del cliente. El alcanzar estos objetivos permitirá a la empresa mejorar sustancialmente sus procesos de segmentación y estrategias comerciales de la empresa.

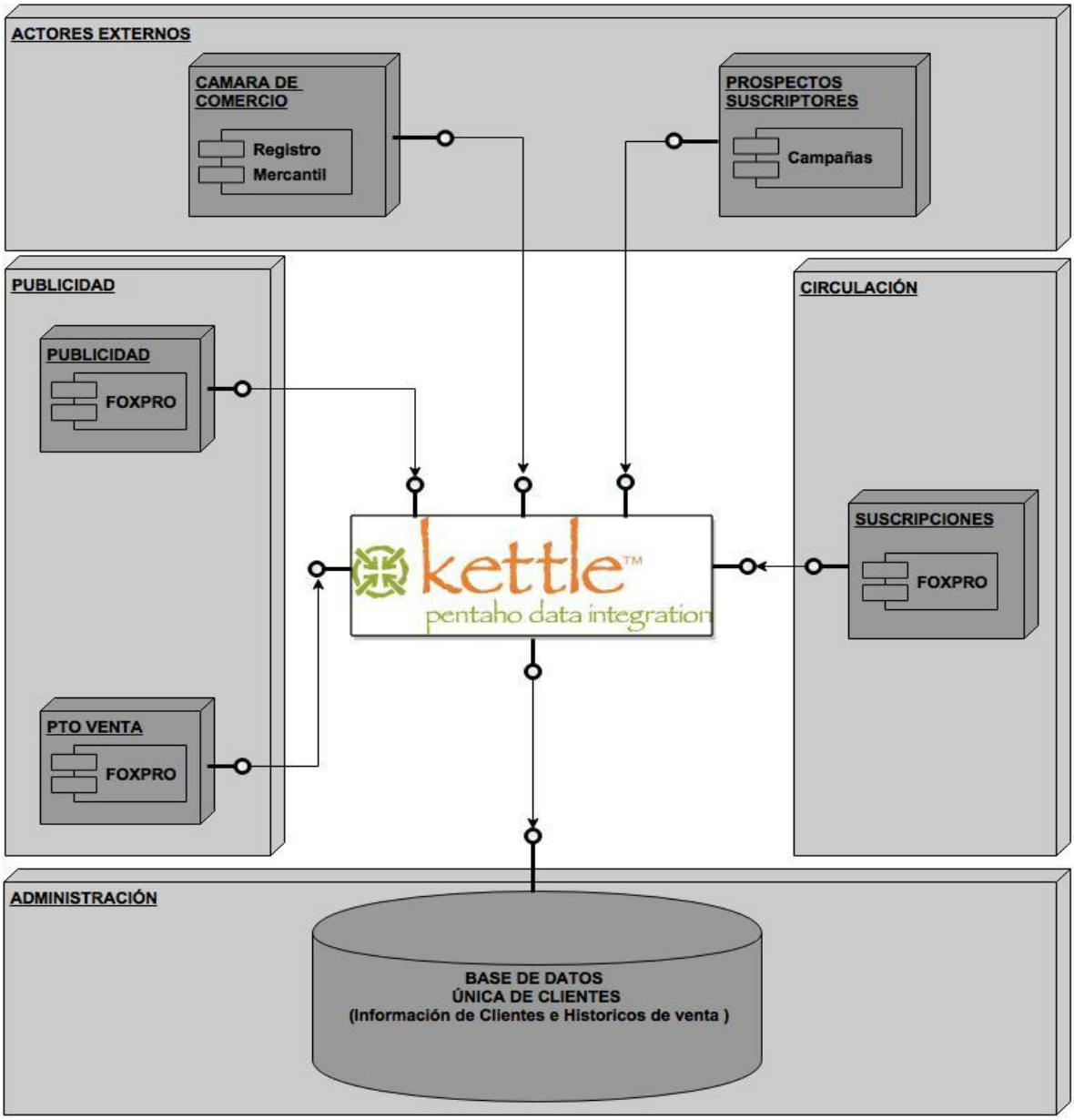
Para entender el modelo anterior es necesario dar una pequeña explicación de los componentes que lo conforman:

Datos externos: Involucra la fuente de datos de la cámara de comercio con un formato de entrada de CSV, que incluye datos como número de identificación, razón social, tipo de sociedad, representante legal entre otros muy valiosos para procesos de validación, gestión y segmentación de la información. La otra fuente incluida en este componente corresponde a la información de prospectos de suscriptores que generalmente es facilitada por convenios establecidos con los socios de negocio de la empresa. Esta información es entregada en formato CSV.

Publicidad: Está compuesto por dos sistemas de información propios desarrollados en Visual FoxPro, que administran la venta de pauta publicitaria y la venta al detal. El formato de su información está basada en tablas libres FoxPro.

Circulación: Conformado por varios sistemas de información propios, pero para efectos del planteamiento de este proyecto solo se toma la información del sistema encargado de administrar los suscriptores del periódico. El formato de su información está basado en tablas libres FoxPro.

Figura 15: Modelo Integrador de Información



Fuente: Autor

- **Administración:** Está definido como el componente que soporta la base de datos única de clientes, la cual tiene como objetivo ser el repositorio central de la información de los clientes además de almacenar los diferentes históricos de las transacciones de cada uno de los sistemas legados de la empresa. Es un repositorio construido en Microsoft SQL.

- Definición de la estructura de datos de la base de datos única de clientes, de acuerdo al análisis de la hoja de vida del cliente. (Datos personales, demográficos, socio económicos, histórico de transacciones, etc.)

- Procedimientos, funciones, consultas y servicios web necesarios para interactuar con la base de datos única de clientes.

- Realizar los desarrollos pertinentes que permitan hacer dependientes los sistemas propios de la empresa, de tal forma que no puedan ingresar ni actualizar información relacionada con clientes de forma directa.

- Tareas de limpieza de datos, estandarización, transformación y validación de la información contra fuentes externas que garanticen la veracidad de la información como la cámara de comercio de Bucaramanga, el RUT (Registro único tributario) y el RUES (Registro único empresarial y social), son consideradas por los encuestados como las más importantes.

- Establecer el tipo de información histórica y forma de migración. Además de definir su periodicidad de migración.

- Definir el plan de unificación de clientes, que permita ir migrando la información normalizada y validada por sistema.

- Establecer un gobierno de datos, que defina las políticas de manejo de la información, que permita definir los procesos y protocolos relacionados con el ingreso de un nuevo cliente y la actualización posterior de su información y completitud de la misma.

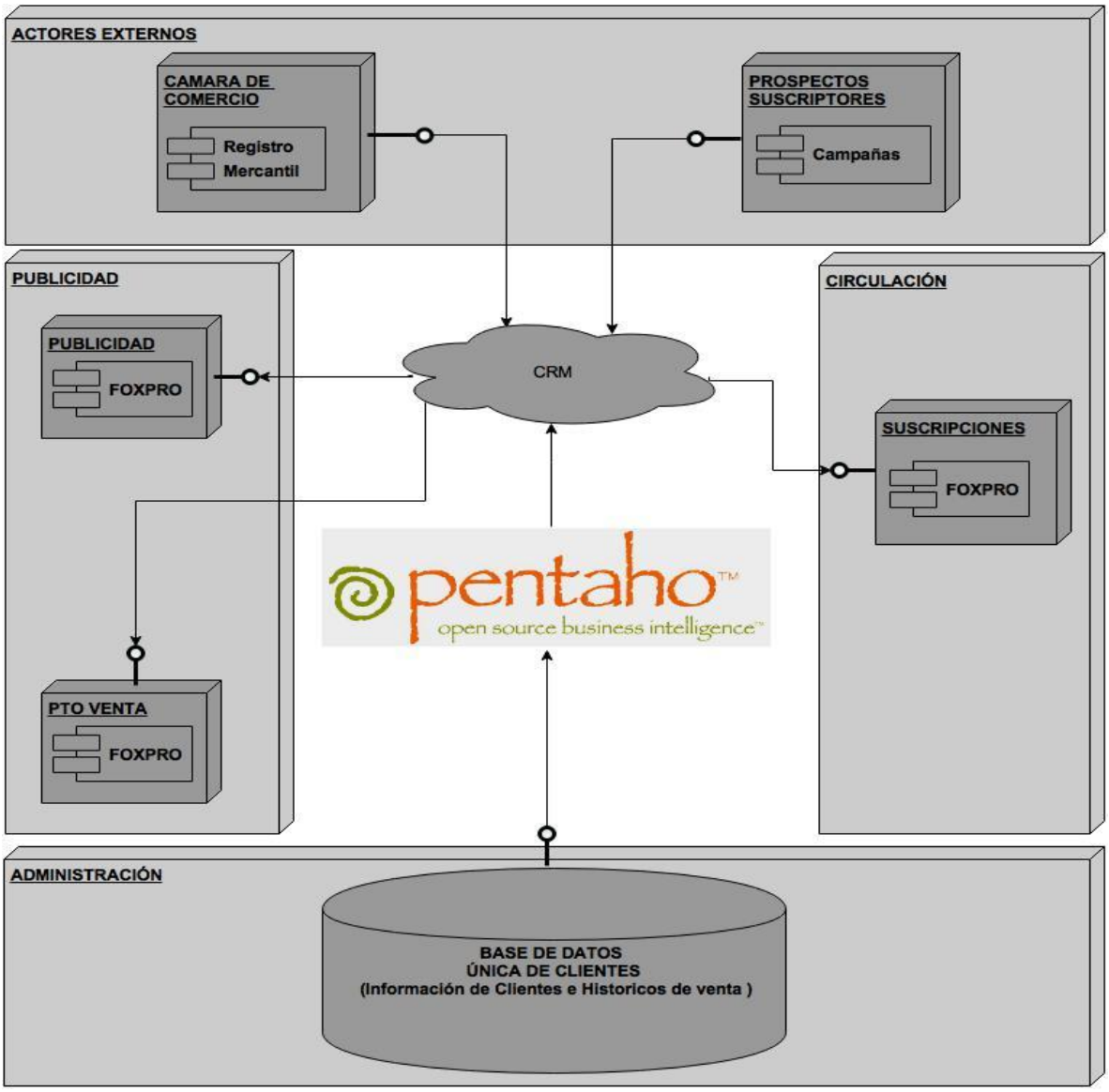
Ahora bien según la arquitectura *Big Data* realizable planteada en la Figura 14. Para las etapas 2 y 3 que tenían que ver con el cargue y la transformación de datos, en este modelo se ve el planteamiento de utilizar la herramienta ETL *Kettle* que está incluida en *Pentaho*, la cual reúne las condiciones necesarias para realizar las tareas planteadas en estas dos etapas.

6.7.2 Modelo de Ventas unificado

La estandarización de los procesos de venta descritos en el capítulo 2 es el objetivo más importante que este modelo persigue. Lo cual permitirá tener un mayor conocimiento del cliente, mediante la implementación de la vista 360 que es

un instrumento que facilita la visualización de los aspectos más relevantes para la fuerza de ventas como: Conocer sus datos básicos, Si es anunciante, Si es Suscriptor, Si es comprador de optativos, de clasificados, etc. A continuación se puede apreciar el modelo. Ver Figura 16.

Figura 16: Modelo de Ventas Unificado



Fuente: Autor

Aunque los componentes de ambos modelos son prácticamente los mismos la funcionalidad que cumplen en cada uno de ellos aporta connotaciones diferentes

en cada caso en particular. Para entender el modelo anterior es necesario dar una pequeña explicación de los componentes que lo conforman:

Datos externos: Involucra la fuente de datos de la cámara de comercio y prospectos de suscriptores ambas fuentes brindan información relevante para la vista 360. Por ejemplo la información de la cámara de comercio proporciona información como la actividad económica muy importante para procesos de segmentación.

Publicidad: Realmente el rol de estos dos sistemas que conforman este componente pasan a ser de ejecutar la orden de publicación y facturación. Su información es retroalimentada a través del modelo integrador.

Circulación: Su funcionalidad en este modelo pasa hacer de soporte a nivel de la administración de la suscripción y de la facturación de la misma. Su información es retroalimentada a través del modelo integrador.

Administración: La base de datos única de clientes en este modelo es fundamental ya que es el ente que agrupa toda la información transaccional e histórica de los sistemas involucrados. Su funcione en este modelo es la de suministrar la información estructurada y unificada de los clientes.

Antes de continuar con las actividades que incluye el modelo, es necesario mencionar un complemento que tiene que ver con la necesidad apremiante de implementar una aplicación de CRM, que administre toda la operación de la venta para las tres unidades de negocio identificadas. Su función principal es coordinar la venta y facturación informando a los sistemas legados de estos eventos.

Este modelo involucra una serie de actividades que sumadas permiten la estandarización de los procesos de venta:

- Canalizar toda gestión de ventas a través del CRM, que le permita a la empresa conocer al cliente, sin intervención del asesor.

- La fuente de datos de la cámara de comercio, pasa hacer una clasificador más de la información suministrada. Por los demás sistemas.

- Lo sistemas que pertenecen a cada uno de los componentes pasan a ser ejecutores de la información de facturación que les provea el CRM.

- La información unificada de los clientes sigue siendo centralizada en la base de datos única de clientes la cual es insumo para la herramienta de *businnes intelligence* que incluye *Pentaho*, que permite el análisis de la información y la visualización

Ahora bien según la arquitectura *Big Data* realizable planteada en la Figura 14. Para la etapa 4 que tiene que ver con la extracción de los datos, en este modelo se aprecia el planteamiento de utilizar la herramienta de *Pentaho BI*, esta herramienta contiene realmente otras más que permiten la analítica y facilidades de visualización que se exigen en esta etapa.

La implementación de los dos modelos están sujetos a una primera fase que involucra unos requerimientos previos que deben darse: Tener en funcionamiento el CRM, haber modificado los sistemas legados para que sean dependientes en lo referente a la creación y modificación de la información de los clientes, tener claras las políticas y protocolos de actualización de información creadas por el gobierno de datos. Una segunda fase que involucra alcanzar el escenario realizable planteado y por ultimo una tercera fase que pretende lograr la implementación del escenario deseado, partiendo de la revaloración de las condiciones. Todo este desarrollo anterior puede ser revisado en el capítulo siguiente.

7. PLAN DE IMPLEMENTACIÓN

En este capítulo se define el plan de implementación de tecnologías *Big Data* para optimizar las estrategias comerciales y de segmentación para Vanguardia Liberal. Se detallan las actividades que conforman el plan y se determina un cronograma determinado por tres fases de implementación, de acuerdo a los dos escenarios planteados en el capítulo 6. Finalmente se discute su viabilidad técnica.

7.1 Actividades

A continuación se especifican tres fases de ejecución de actividades necesarias para poder lograr la implementación de los dos escenarios propuestos en las figuras 15 y 16 del capítulo 6. La primera fase describe los pasos a desarrollar previos al montaje del escenario realizable. La segunda fase plantea las actividades que deberán cumplirse para lograr la puesta en producción del primer escenario. Finalmente la fase tres propone el camino a seguir en el escenario deseable. Las tres fases que se describen involucran una metodología de prototipado que se da durante todo el proceso de desarrollo de la fase, no se describen esos puntos intermedios que se dan para no hacer tan extensa la descripción de las actividades.

Fase 1: Esta fase no forma parte de la propuesta de esta tesis, pero reconociendo su importancia dentro del proceso de implementación se considera dentro de las fases a desarrollar.

- **Desarrollo e Implementación CRM:** El CRM es un componente muy importante ya que se proyecta como el orquestador del modelo de ventas unificado, descrito en la arquitectura planteada (Ver Figura 16), pero aún no forma parte funcional de la misma, debido a que se encuentra en desarrollo. Se estima que este totalmente operativo hacia finales del 2016.

- **Unificación Información de clientes:** Es una actividad que involucra ajustes a los sistemas de publicidad y circulación que eviten el ingreso y modificación de la información del cliente desde cada sistema. Además con lleva todo un proceso validación, verificación de la información de los clientes, esta actividad se viene realizando en el departamento de TI. Debe estar lista para la salida a producción del CRM.

- **Implementación de un Gobierno de Datos:** Este aspecto es fundamental si se quiere entrar en la cultura de tomar decisiones basadas en datos. Su implementación debe darse desde etapas tempranas y mantenerse en el tiempo, mediante el establecimiento de políticas, normas, flujos y protocolos que definan la forma como la información debe actualizarse.

Fase 2: Describe las actividades que permitirán la implementación del escenario realizable planteado en la figura 15 del capítulo 6. El inicio de esta fase depende de tener implementado el CRM, haber unificado los clientes en un solo repositorio central, y tener un gobierno de datos implementado y funcional.

- **Depuración de requerimientos para implementar herramientas de analítica y visualización:** Se deben revisar nuevamente los requerimientos detallados con los usuarios para especificar claramente las fuentes de datos.

- **Depuración de la Arquitectura:** Teniendo en cuenta que la arquitectura planteada esta propuesta a alto nivel, será necesario profundizar más en ella para poder plantearla a nivel de detalle. Esto permitirá identificar componentes necesarios que falten por adquirir o implementar.

- **Instalación de Herramientas y pruebas:** El equipo de TI, debe proceder con la descarga de las herramientas seleccionadas de la página del proveedor y proceder con la instalación correspondiente e iniciar el ciclo de pruebas, que le permitan familiarizarse y adquirir las habilidades necesarias para implementar posteriormente los modelos propuestos.

- **Implementación del Modelo integrador y Modelo de ventas unificado:** Es necesario implementar los dos modelos ya que uno es complemento del otro. El Modelo integrador sería el primero y posteriormente el Modelo Unificado de Ventas. Se adquirirá la tecnología indicada y se realizaran las pruebas y se testeara de acuerdo a los requerimientos planteados.

- **Pruebas de integración y validación de resultados:** Se deben realizar las pruebas de integración de las diferentes fuentes de información especificadas en el modelo integrador, de la misma forma es necesario revisar la consistencia de los datos que propone el modelo de ventas unificado. Bajo el acompañamiento del usuario que garantice los resultados de la implementación de los modelos.

- **Ajustes, capacitación y salida a producción:** Se realizan los últimos ajustes, se prueban y se capacita al usuario para garantizar un proceso limpio en la entrega de los modelos.

- **Mantenimiento y ajustes:** Especifica el tiempo de acompañamiento que puede representar pequeños a justes al modelo que garanticen su correcto funcionamiento.

Fase 3: Esta fase marca el inicio del escenario deseado descrito en el capítulo 6. Se espera que para ese momento sea necesario replantear algunas actividades de las descritas a continuación, basados en que muy seguramente ya existan nuevas necesidades de analítica, visualización, de integración de información tal vez con móviles e incluso se pueda estar pensando en migración a un modelo *cloud*. Es muy probable que siga siendo una realidad, que el negocio continúa sin necesidad de elasticidad en sus procesos de cómputo. Se plantea a nivel de actividades la implementación del prototipo montado en el escenario realizable ahora en el escenario deseado.

- **Depuración de requerimientos de analítica y visualización:** Se deben revisar nuevamente los requerimientos detallados con los usuarios para especificar claramente las fuentes de datos.

- **Depuración de la Arquitectura:** Teniendo en cuenta que la arquitectura planteada esta propuesta a alto nivel, será necesario profundizar más en ella para poder plantearla a nivel de detalle. Esto permitirá identificar componentes necesarios que falten por adquirir o implementar.

- **Instalación de Herramientas y pruebas:** El equipo de TI, debe proceder con la descarga de las herramientas seleccionadas de la página del proveedor y proceder con la instalación correspondiente e iniciar el ciclo de pruebas, que le permitan familiarizarse y adquirir las habilidades necesarias para implementar posteriormente los modelos propuestos.

- **Implementación del Modelo integrador y Modelo de ventas unificado:** Es necesario implementar los dos modelos ya que uno es complemento del otro. El Modelo integrador sería el primero y posteriormente el Modelo Unificado de Ventas. Se adquirirá la tecnología indicada y se realizaran las pruebas y se testeara de acuerdo a los requerimientos planteados.

- **Pruebas de integración y validación de resultados:** Se deben realizar las pruebas de integración de las diferentes fuentes de información especificadas en el modelo integrador, de la misma forma es necesario revisar la consistencia de los datos que propone el modelo de ventas unificado. Bajo el acompañamiento del usuario que garantice los resultados de la implementación de los modelos.

- **Ajustes, capacitación y salida a producción:** Se realizan los últimos ajustes, se prueban y se capacita al usuario para garantizar un proceso limpio en la entrega de los modelos.

- **Mantenimiento y ajustes:** Especifica el tiempo de acompañamiento que puede representar pequeños a justes al modelo que garanticen su correcto funcionamiento.

7.2 Diagrama de Gantt

Una vez determinadas las etapas que se van a realizar durante el plan, es necesario establecer el marco temporal. La Figura 17 muestra un diagrama Gantt con el tiempo de dedicación previsto para cada etapa. El proyecto comienza el 1 de Julio de 2016 y el fin previsto es el 30 de junio de 2018.

Figura 17: Diagrama de Gantt

Fases / Actividades	2016		2017				2018	
	3	4	1	2	3	4	1	2
Fase 1. Componentes Previos								
Desarrollo e Implementación del CRM								
Unificación Información de clientes								
Modificación de sistemas legados								
Implementación gobierno de datos								
Fase 2. Escenario Realizable								
Depuración de requerimientos de analítica y visualización								
Depuración de la arquitectura								
Instalación de herramientas y pruebas								
Implementación Modelo Integrador								
Pruebas de Integración								
Implementación Modelo de ventas unificado								
Pruebas de validación de consistencia								
Ajustes, Capacitación y salida a producción								

Fases / Actividades	2016		2017				2018	
	3	4	1	2	3	4	1	2
Mantenimiento y ajustes								
Fase 3. Escenario Deseado								
Depuración de requerimientos de analítica y visualización								
Depuración de la arquitectura								
Instalación de herramientas y pruebas								
Implementación Modelo Integrador								
Pruebas de Integración								
Implementación Modelo de ventas unificado								
Pruebas de validación de consistencia								
Ajustes, Capacitación y salida a producción								
Mantenimiento y ajustes								

Fuente: Autor

7.3 Discusión viabilidad técnica

Durante el desarrollo del presente trabajo se han analizado los diferentes escenarios frente a la implementación y adopción de tecnologías *Big Data* para la optimización de estrategias comerciales y de segmentación al interior de Vanguardia Liberal.

Este proceso nos ha obligado a cuestionarnos en varios aspectos como: Cual es el mejor tipo de solución frente a las capacidades de infraestructura, habilidades del personal de TI, tiempos de implementación, soporte técnico, presencia en el mercado, comunidad de usuarios, curva de aprendizaje, costo de licenciamiento y escalabilidad. El análisis de estas variables determinaron la selección de la distribución de Cloudera como la herramienta más adecuada para un escenario deseado que involucre a futuro todas las características favorables y necesarias para implementar un esquema 100% *Hadoop* en Vanguardia Liberal. La utilización de Cloudera en este plan se proyecta como viable a partir del tercer trimestre del 2018.

Haber seleccionado Cloudera como la distribución más adecuada en un escenario 100 % *Hadoop*, involucro el reconocimiento de las condiciones actuales de la empresa, frente a demandas de capacidad de procesamiento y almacenamiento. Por lo cual se determinó que Vanguardia Liberal no tiene hoy día unas demandas de elasticidad que nos determine la necesidad de implementar un esquema *Hadoop*. Donde su principal característica son las demandas de capacidad de

cómputo inesperadas. Por tal motivo y teniendo presentes los resultados de la encuesta realizada a Gerentes, Sub Gerentes y Coordinadores de las unidades de negocio de Publicidad, Circulación, Administración y Tecnología, donde se manifiestan las necesidades apremiantes de analítica y visualización. Se determina entonces, diseñar una arquitectura realizable con las tecnologías *Big Data* más adecuadas, basada en las necesidades de analítica y visualización. Teniendo en cuenta lo anterior Vanguardia Liberal podrá tener resultados de la implementación de tecnologías Big Data a un Corto Plazo (9 meses), teniendo plenamente establecida la fase 1 del cronograma, y a un mediano plazo (24 meses), para mostrar que vienen más cosas, incluso la elasticidad.

Después de enfocarnos en una arquitectura realizable, basada en una herramienta que facilitara los procesos de analítica y visualización en forma ágil y a unos costos moderados que no involucraran mayores inversiones a nivel de infraestructura y habiendo evaluado las recomendaciones de Gartner frente a este tipo de herramientas. Se tomó la decisión de implementar la suite de *Pentaho*. Como la opción más robusta a nivel técnico y escalable del mercado.

Teniendo determinada la herramienta, se plantearon dos modelos que permitieran implementar componentes incluidos en la suite de *Pentaho*, como *Kettle* el cual es una herramienta que permite la carga y transformación de datos, Muy necesario en el modelo integrador (Ver Figura 15) y otro modelo en donde se aplican las herramientas de analítica y visualización como es el caso del modelo de ventas unificado en cuyo evento se puede implementar el componente de *Reporting* o *Pivot4J* (Ver Figura 16). Esta plataforma soportará nuevos modelos que en el mediano plazo se identifiquen.

Se elaboró un cronograma de actividades que fue presentado y validado por Tecnología, para determinar el tiempo de implementación de los modelos propuestos, obteniendo como resultado que los dos modelos estarían implementados y en funcionamiento en los dos escenarios en un periodo de 24 meses. Esto indica que la implementación de esta tecnología es completamente viable técnicamente ya que no involucra inicialmente cambios de infraestructura, su costo de implementación a nivel de licencias no aplica por ser *Pentaho* una herramienta *open source*.

8. CONCLUSIONES

- Los procesos de información y las demandas actuales en cuanto a capacidad de cómputo, no exigen tener una elasticidad que haga necesario la implementación de un esquema *Hadoop* en Vanguardia Liberal. Según la literatura la mayor fortaleza de *Big Data* o de soluciones tipo *Hadoop*, es precisamente el poder atender demandas inesperadas y cambios grandísimos en requerimientos de cómputo, que actualmente no se están presentando en la organización.

- La recomendación para iniciar con la adopción de proyectos de *Big Data* es no abordarlo por elasticidad si no por el contrario centrarnos en una necesidad apremiante como la analítica y visualización. reflejadas en la encueta realizada a Gerentes, Sub Gerentes y Coordinadores de Área de las unidades de negocio de Publicidad, Circulación, Administración y Tecnología.

- Basados en las arquitecturas propuestas por el estado del arte de Marcus, Krishnan y Microsoft. Se hace una cuarta propuesta ajustada a Vanguardia Liberal sobre la cual se plantean dos escenarios uno deseable y otro realizable. El deseable plantea la implementación de un ambiente *Hadoop* con la Distribución de Cloudera y el Realizable propone una implementación solo con las herramientas de analítica y visualización que dan solución a las necesidades actuales.

- Basados en el análisis realizado sobre los tipos de solución de *Big Data*, se concluyó que el tipo de solución más adecuado, dadas la condiciones actuales, son las Distribuciones. Recomendando la implementación de la Distribución de Cloudera por ser la más completa y la de mayor proyección a futuro.

- Se concluye que a un mediano plazo se deben volver a evaluar las condiciones de infraestructura, para determinar si una solución tipo *Cloud*, que permite crecer en capacidad de almacenamiento y procesamiento sin la preocupación de la infraestructura de soporte, son viables, ya que podría llegarse a necesitar solo en ciertos eventos puntuales de análisis a futuro.

- Después de haber evaluado las herramientas de análisis y visualización se recomienda la implementación de la *suite* de *Pentaho Comunnity Edition* que incluye componentes de analítica, visualización, minería de datos y procesos ETL entre otros, que satisfacen las necesidades actuales de la organización.

- Se modelaron dos propuestas que dan solución a la problemática actual de islas de información, segmentación, estrategias comerciales y visualización de la vista 360. El modelo integrador da solución a los inconvenientes de islas de información gracias a la integración de la información que ofrece la implementación de las herramientas seleccionadas, de la misma forma soporta la soporta la optimización de los procesos de segmentación y generación de estrategias comerciales. El modelo de ventas unificado soluciona los problemas que impiden la visualización de la vista 360. Al brindar opciones de visualización atreves de cuadros de mandos generados con componentes incluidos en la suite implementada

- Se construyó un plan de implementación de tecnologías *Big Data* correctamente asociado a los procesos de negocio de Vanguardia Liberal, que plantea tres fases la primera representa una etapa de prerrequisitos, la segunda proyecta la el logro de una arquitectura realizable a corto plazo y la tercera fase propone el despliegue de la arquitectura deseable a mediano plazo para un tiempo total de implementación de 24 meses.

- Durante el desarrollo del proyecto se plantean recomendaciones, técnicas y políticas relacionadas con la gestión de la información que soportan los procesos de segmentación y estrategias comerciales. Se indica la necesidad de llevar acabo en una primera fase la implementación de un CRM, la modificación de los sistemas legados y la implementación de un gobierno de datos que fije políticas sobre el manejo de la información y protocolos de actualización. Se propone una arquitectura sin esquema elástico en una segunda fase, se recomienda la implementación de la *suite* de *Pentaho*, para dar solución a los modelos de integración y ventas propuestos. Se proyecta la implementación de la distribución de Cloudera y la posibilidad de migrar a un modelo de nube en una tercera fase

BIBLIOGRAFÍA

5 ventajas de la Inteligencia de Negocios. (s. f.). Recuperado 15 de junio de 2016, a partir de http://mprende.co/gesti%C3%B3n/5-ventajas-de-la-inteligencia-de-negocios

Acerca de Gartner. (s. f.). Recuperado 15 de junio de 2016, a partir de http://www.gartner.com/technology/about.jsp

Acosta Medellín, J. N., & Florez Lara, D. H. (2015). Diseño e implementación de prototipo BI utilizando una herramienta de Big Data para empresas Pymes distribuidoras de tecnología. Recuperado a partir de http://repository.ucatolica.edu.co:8080/xmlui/handle/10983/2543

Apache Mahout: Scalable machine learning and data mining. (s. f.). Recuperado 6 de mayo de 2016, a partir de http://mahout.apache.org/

Apache Spark™ - Lightning-Fast Cluster Computing. (s. f.). Recuperado 21 de febrero de 2016, a partir de https://spark.apache.org/

Arana, E., & Hernan, J. (2015). Plan de marketing para la empresa comercial Jácome", cantón Quevedo, año 2015. Recuperado a partir de http://repositorio.uteq.edu.ec/handle/43000/661

Ardila Cañas, E., & Gómez Díaz, I. C. (2015). Estrategias para la gestión de grandes volúmenes de datos por medio de big data en el contexto de la analítica de negocios: caso MVM ingeniería de software SAS. Recuperado a partir de http://bibliotecadigital.usbcali.edu.co:8080/jspui/handle/10819/2702

AWS | Análisis de Big Data y almacenamiento en la nube. (s. f.). Recuperado 2 de junio de 2016, a partir de //aws.amazon.com/es/big-data/

AWS | Elastic mapreduce (EMR) para el procesamiento rápido de datos. (s. f.). Recuperado 4 de mayo de 2016, a partir de //aws.amazon.com/es/elasticmapreduce/

AWS | Servicio de base de datos gestionada NoSQL (DynamoDB). (s. f.). Recuperado 4 de mayo de 2016, a partir de //aws.amazon.com/es/dynamodb/

AWS | Solución de almacenamiento y análisis de datos en la nube. (s. f.). Recuperado 4 de mayo de 2016, a partir de //aws.amazon.com/es/redshift/

Ayelo, E., & Alberto, S. (2015). Big data como mejora competitiva para la gestión de la información en la agricultura argentina. Recuperado a partir de http://repositorio.udesa.edu.ar/jspui/handle/10908/10919

Big Data Analytics & Hadoop Services from Teradata. (s. f.). Recuperado 7 de mayo de 2016, a partir de http://www.teradata.com.es/services/big-analytics-and-hadoop-services/?ICID=mainnav&LangType=1034&LangSelect=true#tabbable=0&tab1=0&tab2=0

Big Data Appliance | Oracle España. (s. f.). Recuperado 2 de junio de 2016, a partir de https://www.oracle.com/es/engineered-systems/big-data-appliance/index.html

Big data de código abierto para el impaciente, Parte 1: Tutorial Hadoop: Hello World con Java, Pig, Hive, Flume, Fuse, Oozie, y Sqoop con Informix, DB2, y MySQL. (2013, mayo 20). [CT316]. Recuperado 20 de febrero de 2016, a partir de https://www.ibm.com/developerworks/ssa/data/library/techarticle/dm-1209hadoopbigdata/

Cloudera, © 2016, comerciales, I. T. los derechos reservados A. H. y los nombres de los proyectos de código abierto relacionados son marcas comerciales de A. S. F. P. ver una lista completa de las marcas, & Aquí, H. C. (s. f.). Cloudera. Recuperado 2 de junio de 2016, a partir de http://es.cloudera.com/

Cloudera, © 2016, Hadoop, I. A. rights reserved A., trademarks, associated open source project names are trademarks of the A. S. F. F. a complete list of, & Here, C. (s. f.). Why Cloudera. Recuperado 4 de mayo de 2016, a partir de http://es.cloudera.com/

Comparativa y diferencias entre las herramientas de Business Intelligence Pentaho y Qlikview. (s. f.). Recuperado 14 de mayo de 2016, a partir de http://www.buyto.es/general-business-intelligence/comparativa-y-diferencias-entre-pentaho-y-qlikview

Contel Rico, B. (2011). Desarrollo de una solución business intelligence en una empresa del sector de alimentación. Recuperado a partir de https://riunet.upv.es/handle/10251/9127

Coronel, C. (2011). *Bases de Datos, Diseño, Implementacion y Administracion*. Cengage Learning Editores.

Cuadrante Mágico para Data Warehouse y Data Management Solutions para Analytics. (s. f.). Recuperado 6 de mayo de 2016, a partir de https://www.gartner.com/doc/reprints?id=1-2ZSVG83&ct=160229&st=sb&mkt_tok=eyJpIjoiWVRGaFpUbGhNMlk1WVdaaCIsInQiOiJzSktMOWhCSUY3czRcL1Z4N01lTW0zVjJjNlg4aTdaM3hRYk9NN2t4UG9GSmpCRDF0ajJKc253bWY5N3E5NGNmS3NCajFUcWlscHhCTDRBTHdIYWxUSk5NUE41RjB6M3VDNXAwdTNvUktuUEE9In0%253D

Data, C. B. (2015). Conociendo Big Data. *Revista Facultad de Ingeniería (Fac. Ing.)*, *24*(38), 63–77.

Data Visualization & Analytics Software - TIBCO Spotfire. (s. f.). Recuperado 13 de mayo de 2016, a partir de http://spotfire.tibco.com/

Data Warehousing in the Age of Big Data, 1st Edition | Krish Krishnan | ISBN 9780124059207. (s. f.). Recuperado 17 de junio de 2016, a partir de

http://store.elsevier.com/Data-Warehousing-in-the-Age-of-Big-Data/Krish-Krishnan/isbn-9780124059207/

Desarrollo Pentaho - Intryo. (s. f.). Recuperado 14 de mayo de 2016, a partir de http://www.intryo.com/pentaho

Dresner Advisory Services - Home of Business Intelligence and the Wisdom of Crowds ® Market Research. (s. f.). Recuperado 16 de junio de 2016, a partir de http://dresneradvisory.com/

El Proyecto Apache Cassandra. (s. f.). Recuperado 5 de mayo de 2016, a partir de http://cassandra.apache.org/

Ferrer-Sapena, A., & Sánchez-Pérez, E. (2013). Open data, big data:?` hacia dónde nos dirigimos? *Anuario ThinkEPI 2013*, *7*, 150–156.

Ferri, C., Ramírez, Q. M. J., & Hernández, O. J. (2004). Introducción a la minería de datos. *Editorial Prentice Hall, España*.

Forrester : Welcome. (s. f.). Recuperado 16 de junio de 2016, a partir de https://www.forrester.com/home/

Garcés Uquillas, M. B. (2015). Estudio comparativo de metodologías e implementación de alternativas business intelligence opensource vs. propietarias en entornos tradicionales; caso prototipo en las pymes en el sector agroindustrial. Recuperado a partir de http://dspace.udla.edu.ec/handle/33000/2660

García, J. H. M. (2010). La Inteligencia De Negocios Como Herramienta Para La Toma De Decisiones Estratégicas En Las Empresas. Análisis De Su Aplicabilidad En El Contexto Corporativo Colombiano. Recuperado a partir de http://www.docentes.unal.edu.co/hrumana/docs/TESIS_JHMG_Inteligencia_de_Negocios_2010.pdf

Gartner BI. (s. f.). Recuperado 6 de mayo de 2016, a partir de https://www.gartner.com/doc/reprints?id=1-2XXET8P&ct=160204&st=sb

Google. (s. f.). What is BigQuery? Recuperado 4 de mayo de 2016, a partir de https://cloud.google.com/bigquery/what-is-bigquery

Goyzueta Rivera, S. I. (2015). Big Data Marketing: una aproximación. *Revista Perspectivas*, (35), 147–158.

Guerrero López, F. A., Rodríguez Pinilla, J. E., & others. (2014). Diseño y desarrollo de una guía para la implementación de un ambiente Big Data en la Universidad Católica de Colombia. Recuperado a partir de http://repository.ucatolica.edu.co/jspui/handle/10983/1320

Guevara, S., & Antonio, O. (2015). Modelo de inteligencia de negocio para la toma de decisiones en la empresa San Roque S.A. *Tesis digitales - UPAO*. Recuperado a partir de http://repositorio.upao.edu.pe/handle/upaorep/794

Guided Analytics | Business Intelligence Software | QlikView. (s. f.). Recuperado 5 de mayo de 2016, a partir de http://www.qlik.com/products/qlikview

Gutiérrez, J. A. T., Acebrón, L. B., & Casielles, R. V. (2005). *Investigación de mercados: métodos de recogida y análisis de la información para la toma de decisiones en marketing*. Editorial Paraninfo.

Hadoop - IBM - Apache Hadoop Open Source Software Project. (2016-05-04). Recuperado 4 de mayo de 2016, a partir de http://www.ibm.com/analytics/us/en/technology/hadoop/

Hadoop Summit: la seguridad un reto de Big Data según Hortonworks. (s. f.). Recuperado 6 de mayo de 2016, a partir de http://www.revistabyte.es/actualidad-byte/seguridad-principal-reto-de-big-data/

Hortonworks: Open and Connected Data Platforms. (s. f.). Recuperado 2 de junio de 2016, a partir de http://hortonworks.com/

IBM - Colombia. (2016, mayo 16). Recuperado 16 de junio de 2016, a partir de https://www.ibm.com/co-es/

IBM adquiere Cloudant. (2014, febrero 24). [CTB10]. Recuperado 5 de mayo de 2016, a partir de https://www-03.ibm.com/press/mx/es/pressrelease/43317.wss

IBM - InfoSphere Information Server - Data Integration, Information Integration - Overview. (s. f.). Recuperado 4 de mayo de 2016, a partir de http://www-01.ibm.com/software/data/integration/info_server/

IBM Analytics - Stream Computing. (2015, julio 3). [ct000]. Recuperado 4 de mayo de 2016, a partir de http://www.ibm.com/analytics/us/en/technology/stream-computing/

IBM DB2 for Linux, Unix and Windows – Database software – IBM Analytics. (2016, marzo 17). Recuperado 4 de mayo de 2016, a partir de https://www.ibm.com/analytics/us/en/technology/db2/db2-linux-unix-windows.html

IBM PureData System - Analytics - System. (s. f.). Recuperado 4 de mayo de 2016, a partir de http://www-01.ibm.com/software/data/puredata/analytics/index.html

IBM SPSS - IBM Analytics. (2016, marzo 17). Recuperado 6 de mayo de 2016, a partir de http://www.ibm.com/analytics/us/en/technology/spss/

IBM Watson Explorer. (s. f.). [CT004]. Recuperado 4 de mayo de 2016, a partir de http://www.ibm.com/smarterplanet/us/en/ibmwatson/explorer.html

Inteligencia Operacional, Administración de registros, Administración de aplicaciones, Seguridad y cumplimiento de empresa. (s. f.). Recuperado 4 de mayo de 2016, a partir de http://www.splunk.com/es_es

Is Big Data Still a Thing? (The 2016 Big Data Landscape). (2016, febrero 1). Recuperado 2 de mayo de 2016, a partir de http://mattturck.com/2016/02/01/big-data-landscape/

Las bases de datos NoSQL. (s. f.). Recuperado 15 de junio de 2016, a partir de http://nosql-database.org/

Lawrence, P. S. (2002). Ingeniería de Software, Teoría y Práctica. *Editorial Prentice Hall. Primera edición. ISBN*, 987–9460.

López García, D. (2013). Análisis de las posiblidades de uso de Big Data en las organizaciones. Recuperado a partir de http://repositorio.unican.es/xmlui/handle/10902/4528

MapR: Plataforma de datos convergente. (s. f.). Recuperado 5 de mayo de 2016, a partir de https://www.mapr.com/

Más de 500 mil usuarios únicos de 45 mil empresas en 185 países ayudaron a darle forma al nuevo Power BI | News Center Latinoamérica. (s. f.). Recuperado a partir de https://news.microsoft.com/es-xl/mas-de-500-mil-usuarios-unicos-de-45-mil-empresas-en-185-paises-ayudaron-a-darle-forma-al-nuevo-power-bi/#sm.001k62cxz12ptdj1thb19crn2ff24

Microsoft: página principal. (s. f.). Recuperado 16 de junio de 2016, a partir de https://www.microsoft.com/es-co/

MLlib | Spark Apache. (s. f.). Recuperado 6 de mayo de 2016, a partir de http://spark.apache.org/mllib/

Morales, G., & Carolina, S. (2015). Estudio Comparativo de Métodos Existentes para Integrar la Información Estructurada y no Estructurada de una Industria Enfocado en la Generación de Conocimiento, Desde la Perspectiva de una Solución Integral de Big Data. Recuperado a partir de http://dspace.udla.edu.ec/handle/33000/3385

Pentaho | Data Integration and Business Analytics Platform for Big Data Deployments. (s. f.). Recuperado 5 de mayo de 2016, a partir de http://www.pentaho.com/homepage/homepage

Power BI | Herramientas de BI para la visualización de datos interactivos. (s. f.). Recuperado 13 de mayo de 2016, a partir de https://powerbi.microsoft.com/es-es/

PSPP. (s. f.). Recuperado 6 de mayo de 2016, a partir de http://www.gnu.org/software/pspp/

Reinventando la gestión de datos. (s. f.). Recuperado 16 de junio de 2016, a partir de https://www.mongodb.com/dynamic/node

Rodriguez, J. S., & Valverde, E. L. (s. f.). Big Data Analytics: propuesta de una arquitectura. Recuperado a partir de http://bb9.ulacit.ac.cr/tesinas/publicaciones/043235.pdf

Romero Albarracín, D. L., Vargas López, C. A., Rojas Cordero, A., & Director. (2016, febrero 25). *Diseño de prototipo para la implementación de un sistema Big Data* (Thesis). Recuperado a partir de http://alejandria.poligran.edu.co/handle/10823/788

Sabater Picañol, J. (2013). Big Data. Recuperado a partir de http://upcommons.upc.edu/handle/2099.1/20144

SAS. (s. f.). Recuperado 16 de junio de 2016, a partir de http://www.sas.com/en_us/insights/big-data.html

SelectHub. (2016, enero 6). Tableau vs QlikView vs Microsoft Power BI. Recuperado a partir de https://selecthub.com/business-intelligence/tableau-vs-qlikview-vs-microsoft-power-bi/

Serrat Morros, R. (2013). Big Data: análisis de herramientas y soluciones. Recuperado a partir de http://upcommons.upc.edu/handle/2099.1/19855

Silva Guerra, H. (2014). LOS EFECTOS DE LA IMAGEN, LOS SÍMBOLOS Y LOS HÁBITOS CULTURALES EN LA ACTITUD CONSUMISTA DEL NEGOCIO MINORISTA COLOMBIANO (Spanish). *University of St. Gallen, Business Dissertations*, 1-160.

Software, P. (2015, febrero 13). Pivotal [text/html]. Recuperado 6 de mayo de 2016, a partir de http://pivotal.io/

Soluciones de plataforma de análisis de big data en la nube – Haven OnDemand | HP® Colombia. (s. f.). Recuperado 4 de mayo de 2016, a partir de http://www8.hp.com/co/es/software-solutions/big-data-cloud-haven-ondemand/index.html

Tableau Software. (s. f.). Recuperado 6 de mayo de 2016, a partir de http://www.tableau.com/

TensorFlow - una biblioteca de software de fuente abierta para la máquina de Inteligencia. (s. f.). Recuperado 27 de mayo de 2016, a partir de https://www.tensorflow.org/

The BI Survey 15, BARC's annual report on the BI industry. (s. f.). Recuperado a partir de http://barc-research.com/bi-survey-15/

Un nuevo punto de Vista sobre Inteligencia de Negocios – Microsoft Power BI. (s. f.). Recuperado 14 de mayo de 2016, a partir de http://www.icsicorp.com.mx/index.php/news/113-un-nuevo-punto-de-vista-sobre-inteligencia-de-negocios-business-intelligence-bi-microsoft-power-bi

Vega, J. J. C., Ortega, J. F. C., & Aguilar, L. J. (2015). Arquitectura Tecnológica Para Big Data. *Revista Científica*, *21*. Recuperado a partir de http://revistas.udistrital.edu.co/ojs/index.php/revcie/article/view/8451

Virtualización de servidor con VMware vSphere | VMware Colombia. (s. f.). Recuperado 4 de mayo de 2016, a partir de http://www.vmware.com/co/products/vsphere/

Vivancos Vicente, P. J. (2016, enero 20). Plataforma inteligente de diseño para todos para control de teléfonos móviles mediante habla en lenguaje natural [info:eu-repo/semantics/doctoralThesis]. Recuperado 21 de febrero de 2016, a partir de https://digitum.um.es/xmlui/handle/10201/47541

Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (s. f.). Recuperado 6 de mayo de 2016, a partir de http://www.cs.waikato.ac.nz/ml/weka/

Welcome to Apache™ Hadoop®! (s. f.). Recuperado 5 de mayo de 2016, a partir de https://hadoop.apache.org/

Zeppelin. (s. f.). Recuperado 6 de mayo de 2016, a partir de https://zeppelin.incubator.apache.org/

ANEXOS

ANEXO A. ENCUESTA

Propósito:

En el marco de la Maestría en Gestión Aplicación y Desarrollo de Software de la Universidad Autónoma de Bucaramanga, el profesional Francisco Carrillo Álvarez se encuentran desarrollando su proyecto de grado, que tiene como propósito realizar una investigación de tipo cualitativa, donde se emplea el método de estudio de caso y la unidad de análisis es el periódico Vanguardia Liberal, con el objetivo de plantear un plan de implementación de tecnologías *Big Data* para la optimización de estrategias comerciales y de segmentación.

Los resultados que se obtengan en esta investigación, Vanguardia Liberal podrá recibirlos como insumos para tenerlos en cuenta en proyectos que desarrolle en el contexto de la analítica de negocios. Por último, la información que se obtenga será manejada bajo estricta confidencialidad y solo tendrá fines académicos.

Objetivo:

El objetivo de este instrumento de recolección de información (tipo encuesta) es conocer la opinión de Gerentes, Sub Gerentes y Coordinadores de área de las unidades de negocio de Circulación, Publicidad, Administración y Tecnología de Vanguardia Liberal en el contexto de la gestión de la información y analítica de negocios, con el fin de recibir sus contribuciones en relación a la gestión de grandes volúmenes de información estructurada y desestructurada para la implementación de soluciones y/o proyectos de *Big Data*.

Preguntas:

Anexo 1. Encabezado de la encuesta

1. ¿Qué tan importante considera los siguientes aspectos para segmentar su negocio? Al contestar tenga en cuenta la siguiente escala: 1. Nada Importante, 2. Un poco Importante 3. Muy Importante 4. Extremadamente Importante *

	1	2	3	4
Tipo persona (Natural o Jurídica)	○	○	○	○
Fecha de Nacimiento/Constitución	○	○	○	○
Genero	○	○	○	○
Estrato	○	○	○	○
Ubicación Geográfica (Dirección, Barrio, Comuna)	○	○	○	○
Código Postal	○	○	○	○
Actividad Económica	○	○	○	○
Tamaño de la Empresa (Pequeña, Mediana, Grande, Micro)	○	○	○	○
Nivel de Ingresos/Volumen de Ventas	○	○	○	○
Número de empleados	○	○	○	○
Página Web (Número de Visitantes)	○	○	○	○
Twitter (Número de Seguidores)	○	○	○	○
Facebook (Numero de Likes)	○	○	○	○
Instagram (Número de Conexiones)	○	○	○	○
LinkedIn (Número de Conexiones)	○	○	○	○

Anexo 2. Encuesta pregunta 1

Monto Compras Último Mes	○	○	○	○
Monto Compras Ultimo año	○	○	○	○
Frecuencia de compras	○	○	○	○
Ocupación	○	○	○	○
Hobbies	○	○	○	○
Número de hijos	○	○	○	○
Estado civil	○	○	○	○
Nivel de Escolaridad	○	○	○	○
Profesión	○	○	○	○
Pensionado (Si/No)	○	○	○	○

2. ¿Cuáles de las siguiente fuentes externas de información consideraría importantes incluir dentro de su proceso de negocio? Marque con una (X) *

- [ ] Cámara de Comercio
- [ ] RUNT (Registro único de Transito)
- [ ] FOSYGA (Fondo de Solidaridad y Garantía. Ministerio de Salud)
- [ ] RUT (Registro Único Tributario. DIAN)
- [ ] RUAF (Registro Único de Afiliados. Ministerio de Salud)
- [ ] RUES (Registro único empresarial y social - Cámaras de Comercio)
- [ ] Aliados Comerciales
- [ ] Otro: _____

Anexo 3. Encuesta pregunta 2

3. Califique de acuerdo al nivel de relevancia la información interna con la cual le gustaría contar en la gestión de su proceso de negocio. Al contestar tenga en cuenta la siguiente escala: 1.Nada Relevante 2.Un poco Relevante 3.Muy Relevante 4.Extremadamente Relevante *

	1	2	3	4
Es cliente de pauta publicitaria	○	○	○	○
Es suscriptor del periódico	○	○	○	○
Es cliente de optativos	○	○	○	○
Es cliente de clasificados	○	○	○	○
Es proveedor	○	○	○	○
Monto pauta publicitaria último periodo	○	○	○	○
Monto de optativos último periodo	○	○	○	○
Monto clasificados último periodo	○	○	○	○
Desde cuando es suscriptor	○	○	○	○
Está vigente su suscripción	○	○	○	○
Cuando se le vence su suscripción	○	○	○	○
Qué tipo de suscripción tiene	○	○	○	○
Ultimo asesor de publicidad	○	○	○	○
Ultimo asesor de circulación	○	○	○	○
Ultimo optativo que adquirió	○	○	○	○

Anexo 4. Encuesta pregunta 3

Fecha de clasificado más reciente	○	○	○	○
Fecha de publicación aviso más reciente	○	○	○	○

4 ¿Qué nivel de importancia tendría para Ud., los siguientes aspectos para obtener una analítica de datos masivos que brinde un mejor apoyo en la gestión de su proceso?. Al contestar tenga en cuenta la siguiente escala: 1.Nada Importante 2.Un poco Importante 3.Muy Importante 4.Extremadamente Importante *

	1	2	3	4
Incluir más Fuentes de datos externas	○	○	○	○
Incluir más Fuentes de datos internas	○	○	○	○
Retirar algunas fuentes de datos	○	○	○	○
Preparación de datos, carga automatizada e integración de los datos dentro de un flujo de proceso.	○	○	○	○
Almacenamiento para posterior análisis y toma de decisiones (Análisis histórico de la dinámica de los clientes)	○	○	○	○
Análisis de la Información y Generación de Conocimiento, para la toma de decisiones	○	○	○	○

5. Basado en su interés y de acuerdo al proceso de negocio que gestiona, determine el nivel de impacto de los siguientes beneficios empresariales derivados del uso y análisis masivo de mayores fuentes de datos internas y externas. Al contestar tenga en cuenta la siguiente escala: 1.Nada Impactante 2.Un poco Impactante 3.Muy Impactante 4.Extremadamente Impactante *

Anexo 5. Encuesta pregunta 4

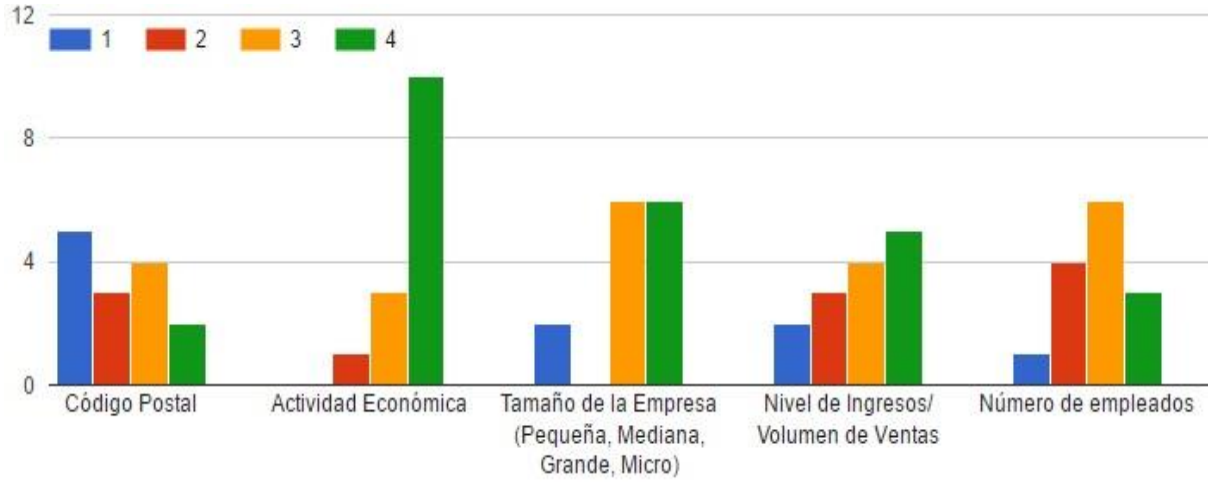
	1	2	3	4
Nuevas Fuentes de Ingresos	○	○	○	○
Aumento de índices de retención y adquisición de clientes	○	○	○	○
Desarrollo de nuevos productos y servicios	○	○	○	○
Mejora de la Experiencia de Cliente	○	○	○	○
Mejora de la competitividad	○	○	○	○

Anexo 6. Encuesta pregunta 2

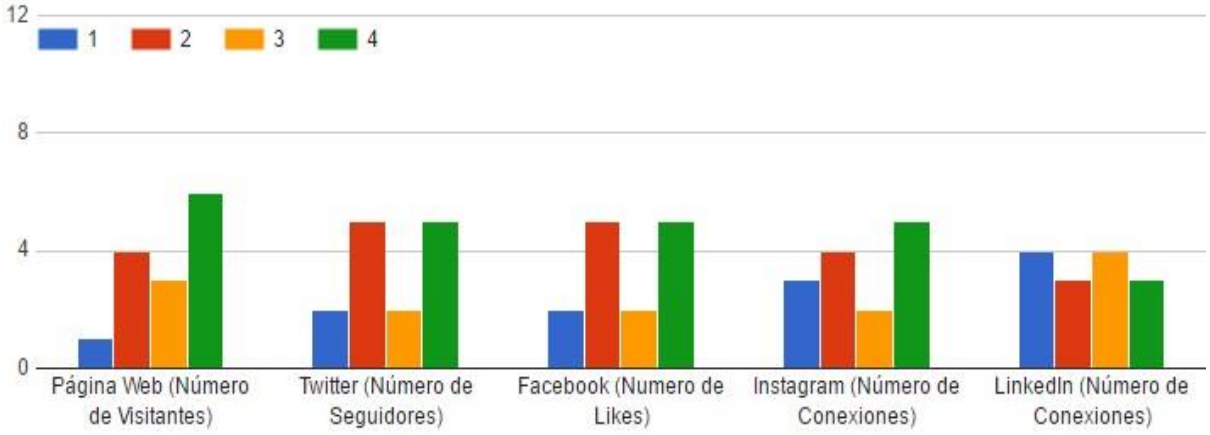
ANEXO B. RESULTADOS ENCUESTA

A continuación se detalla el resultado obtenido por cada una de las preguntas realizadas en el instrumento de recolección de información (encuesta).

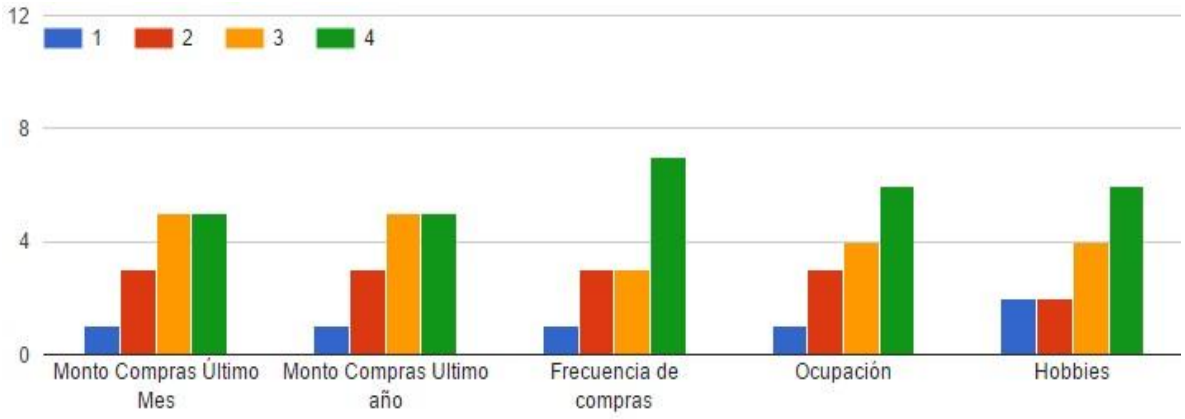
Anexo 7. Resultado pregunta 1 parte A de la encuesta (Google Drive)



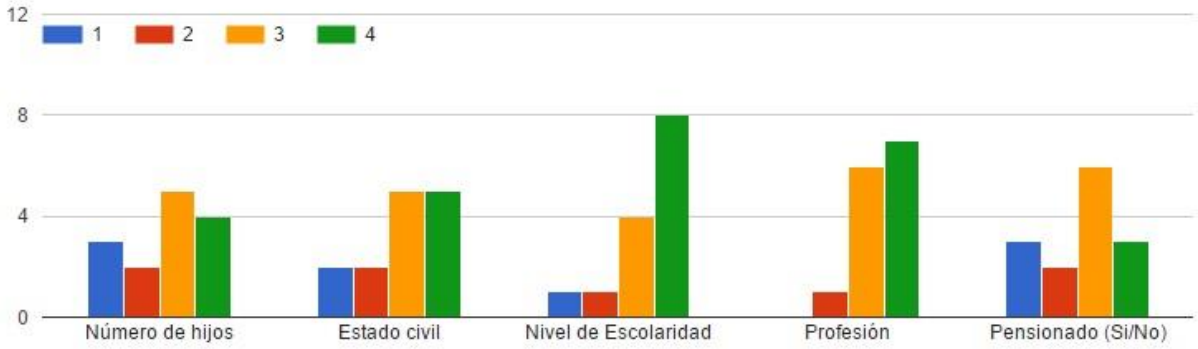
Anexo 8. Resultado pregunta 1 parte B de la encuesta (Google Drive)



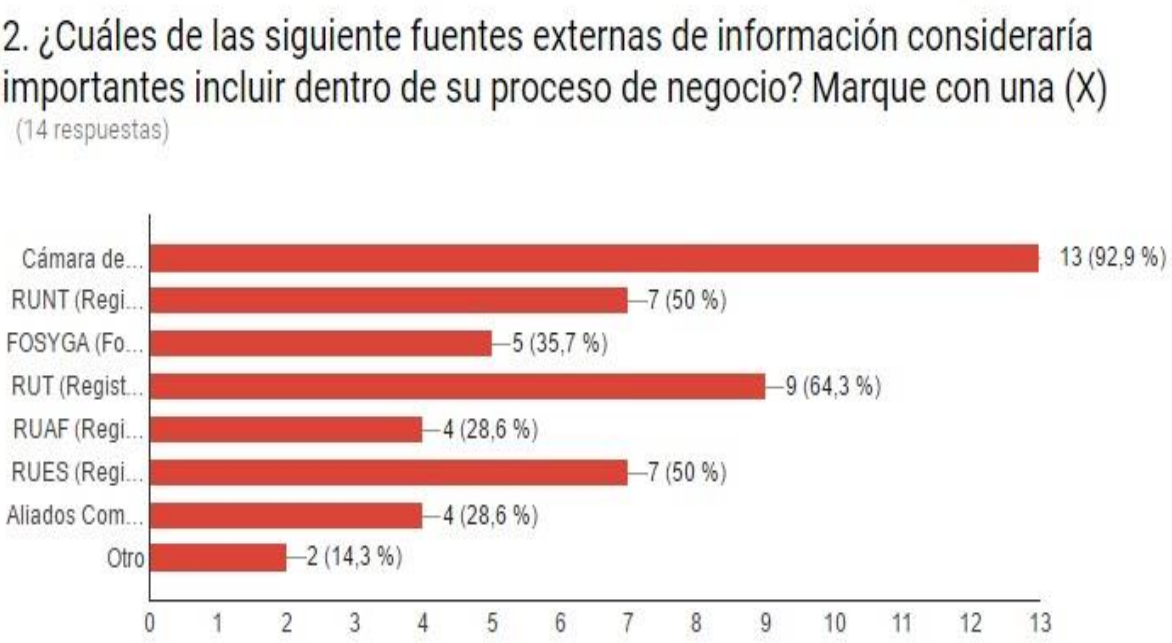
Anexo 9. Resultado pregunta 1 parte C de la encuesta (Google Drive)



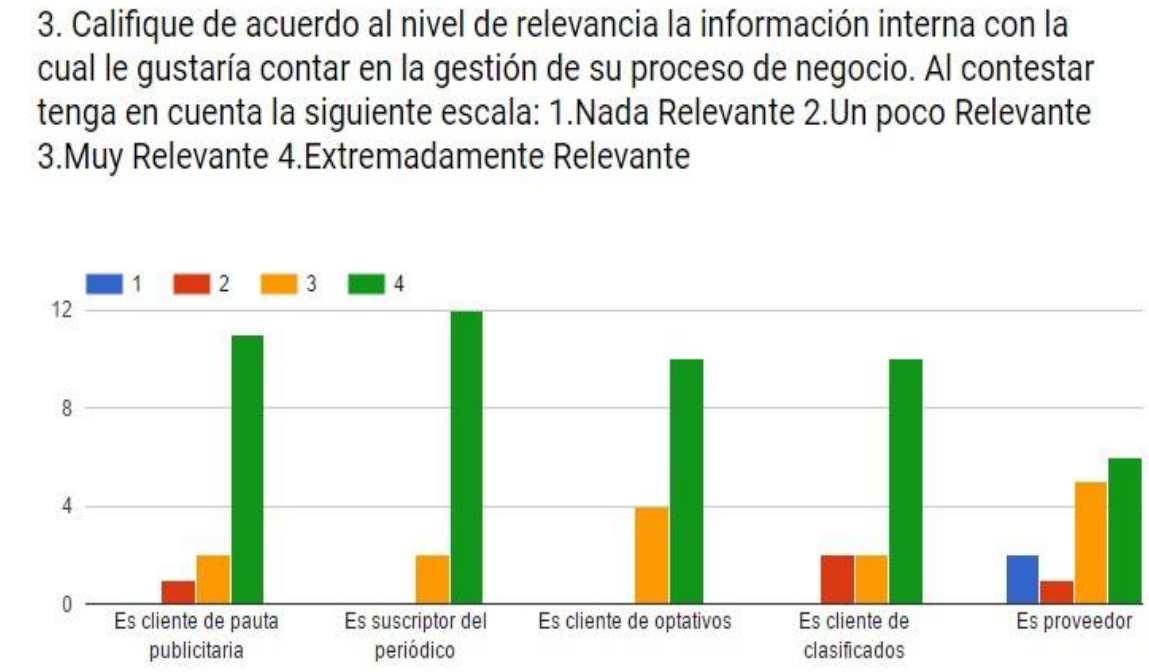
Anexo 10. Resultado pregunta 1 parte D de la encuesta (Google Drive)



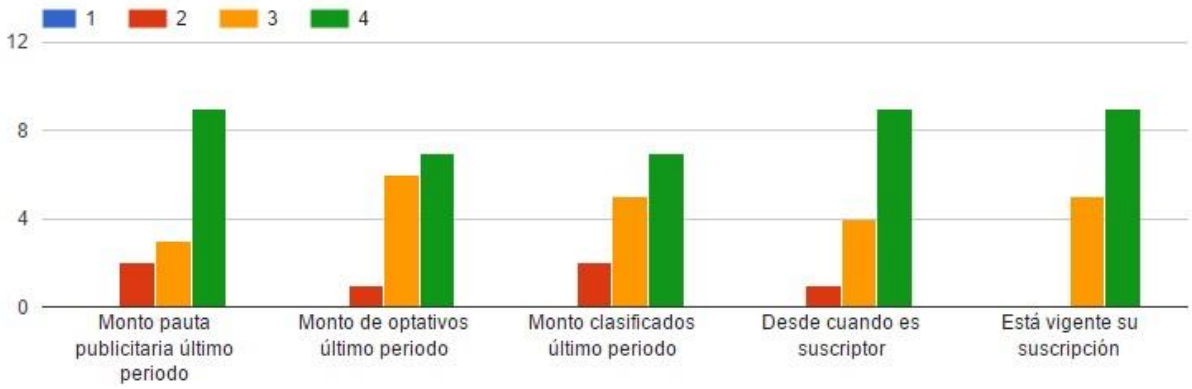
Anexo 11. Resultado pregunta 1 parte E de la encuesta (Google Drive)



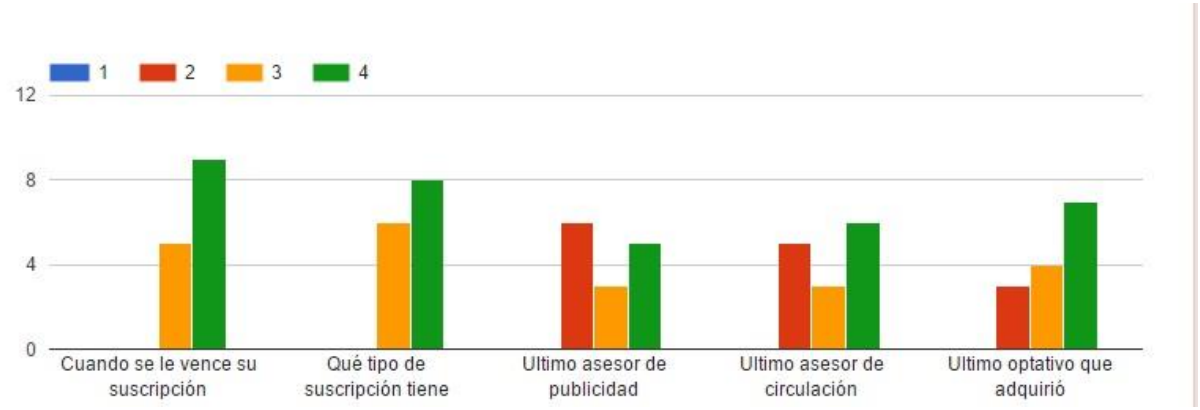
Anexo 12. Resultado pregunta 2 de la encuesta (Google Drive)



Anexo 13. Resultado pregunta 3 parte A de la encuesta (Google Drive)



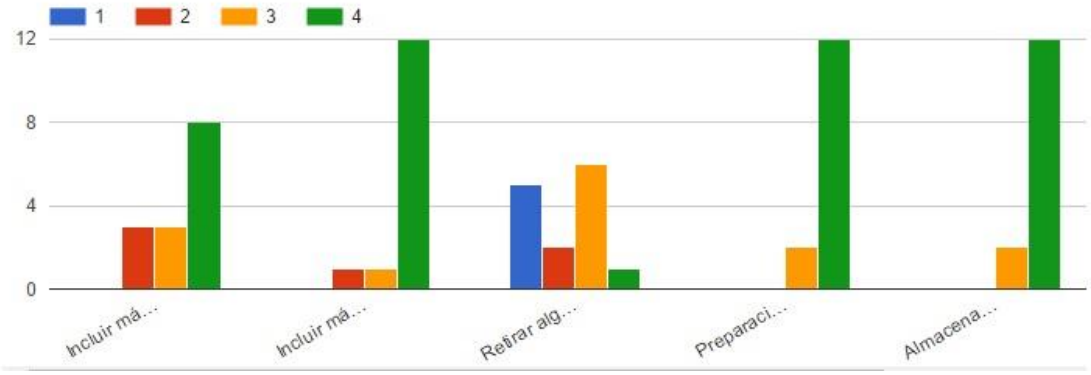
Anexo 14. Resultado pregunta 3 parte B de la encuesta (Google Drive)



Anexo 15. Resultado pregunta 3 parte C de la encuesta (Google Drive)

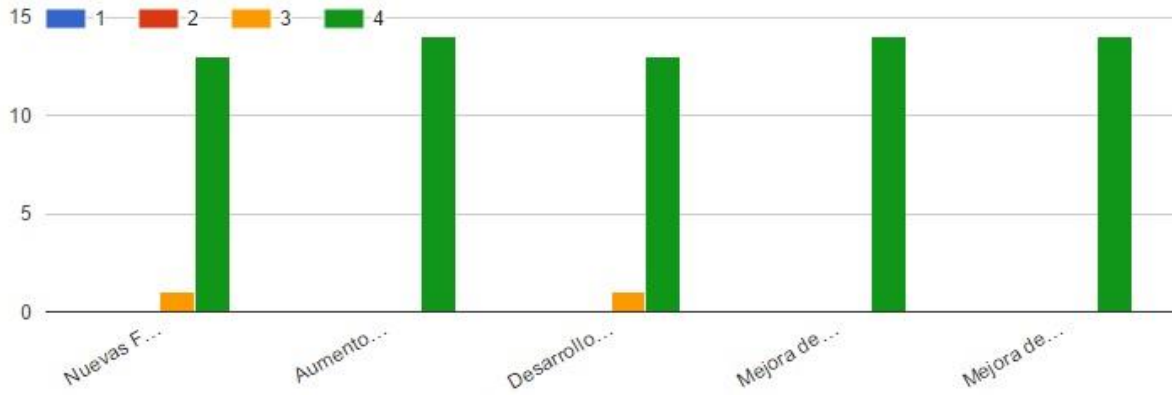
Anexo 16. Resultado pregunta 3 parte D de la encuesta (Google Drive)

4 ¿Qué nivel de importancia tendría para Ud., los siguientes aspectos para obtener una analítica de datos masivos que brinde un mejor apoyo en la gestión de su proceso?. Al contestar tenga en cuenta la siguiente escala: 1.Nada Importante 2.Un poco Importante 3.Muy Importante 4.Extremadamente Importante



Anexo 17. Resultado pregunta 4 de la encuesta (Google Drive)

5. Basado en su interés y de acuerdo al proceso de negocio que gestiona, determine el nivel de impacto de los siguientes beneficios empresariales derivados del uso y análisis masivo de mayores fuentes de datos internas y externas. Al contestar tenga en cuenta la siguiente escala: 1.Nada Impactante 2.Un poco Impactante 3.Muy Impactante 4.Extremadamente Impactante



Anexo 18. Resultado pregunta 5 de la encuesta (Google Drive)

ANEXO C. DESCRIPCIÓN HERRAMIENTAS

A continuación se hace una descripción breve de las herramientas que presentan una mejor viabilidad de adopción, teniendo en cuenta aspectos como el respaldo ofrecido a nivel de soporte por su casa matriz, evolución tecnológica, utilidades ofrecidas, implementación del ecosistema *hadoop* dentro de sus componentes, facilidad de uso, escalabilidad, curva de aprendizaje, costo de licenciamiento entre otros:

Hadoop

Inspirado inicialmente por trabajos publicados por Google, se ha convertido en el estándar para el almacenamiento, procesamiento y análisis de datos. Es un *framework* de código abierto 100% que pertenece a la fundación apache, fue pionero en una nueva forma de almacenamiento y procesamiento de información. *Hadoop* permite la distribución del procesamiento en paralelo de grandes cantidades de datos a través de servidores de bajo costo utilizando estándares de la industria y pudiendo escalar sin límites.

Maneja todo tipo de fuentes de datos: estructurado, no estructurado, archivos, imágenes, audio, correo electrónico y casi cualquier cosa independiente de su formato nativo.

Una de las ventajas en costos es que se basa en una estructura interna de datos que es posible implementar en servidores estándar de la industria, en lugar de sistemas de almacenamiento de datos especializados.(Ayelo & Alberto, 2015)

Soluciones IBM para *Big Data*

Entre las tecnologías ofrecidas por esta compañía se incluyen hardware de servidor y de almacenamiento, software de base de datos, aplicaciones analíticas y servicios asociados. Los productos más conocidos son:

***InfoSphere Streams*:** Permite el análisis continuo de volúmenes masivos de *streaming* con tiempos de respuesta inferiores a un milisegundo.(«IBM *Analytics - Stream Computing*», 2015)

***InfoSphere BigInsights*:** Solución basada en *Hadoop*, lista para la suplir las necesidades de la empresa y el análisis de grandes volúmenes de datos

estructurados y no estructurados.(«*Hadoop* - IBM - *Apache Hadoop Open Source Software Project*», 2016-05-04)

IBM Watson Explorador: Proporciona búsqueda, navegación y descubrimiento sobre una amplia gama de fuentes de datos y aplicaciones, tanto dentro como fuera de la empresa para ayudarle a descubrir la información y los conocimientos más relevantes.(«IBM Watson Explorer», s. f.)

IBM PureData (Impulsado por la tecnología Netezza): Simplifica y optimiza el rendimiento de los servicios de datos para aplicaciones analíticas, permitiendo correr algoritmos muy complejos con resultados en minutos.(«IBM *PureData System - Analytics - System*», s. f.)

DB2 con BLU Aceleración: Capacidad avanzada de innovación para acelerar el trabajo analítico en bases de datos y *Data WareHouse*. («IBM DB2 for Linux, Unix and Windows – Database software – IBM Analytics», 2016)

***InfoSphere Information Server*:** Permite limpiar y transformar datos, para luego entregar información confiable a la empresa o negocio. Esta herramienta permite trabajar inteligencia de negocios, facilitando la mejor toma de decisiones.(«IBM - *InfoSphere Information Server - Data Integration, Information Integration - Overview*», s. f.)

IBM dentro de sus funcionalidades ha adoptado el uso de la plataforma de análisis de datos de código abierto *Hadoop*.

HP Haven

Ofrece una mezcla de hardware, software y servicios. Es conocida por la plataforma de análisis Vertica. HP lleva su plataforma de *Big Data* a la nube con *HP Haven OnDemand* ofrece una ruta rápida hacia la información basada en datos. Se trata de un paquete de productos de servicios en la nube de fácil consumo que tendrá instalado y en funcionamiento en cuestión de minutos.

Las API orientadas a los desarrolladores y los análisis avanzados le permiten crear con facilidad servicios y aplicaciones. Incluye una colección líder en el sector de más de 50 API con búsqueda contextual reconocimiento de voz y facial para incorporar la compresión humana a los datos sin estructurar en todas las formas, desde el texto libre a la transmisión de video y el internet de las

cosas.(«Soluciones de plataforma de análisis de big data en la nube – Haven OnDemand | HP® Colombia», s. f.)

Teradata

Teradata es líder mundial reconocido en el desarrollo de soluciones analíticas y la innovación de *Data Warehouse*. Ofrece portafolio para *Hadoop*, servicios e integración de *Big Data*.

La compañía de investigación de marketing Gartner Group ubico Teradata en el cuadrante de líderes en los años 2009, 2010 y 2012 en sus informes "*Magic Quadrant for DataBase Data Warehouse Management Systems*". TeraData comenzó asociarse con el término "*Big Data*" en el año 2010 con el surgimiento de nuevos medios de comunicación, como las redes sociales.

El aumento en datos semi-estructurados y no estructurados obtenidos de interacciones en línea impulsó a TeraData para formar el "Club Petabyte" en el 2011 para sus usuarios de grandes datos.(Ayelo & Alberto, 2015)

Oracle Fast Data Solutions

Está diseñado para optimizar la eficiencia y la escala en temas de procesamiento y transacciones de gran volumen.

***Oracle Event Processing*:** Filtra y correlaciona datos utilizando reglas predefinidas a través de las fuentes de grandes cantidades de datos. Cuando se ejecuta con *Oracle Coherence*, se ejecuta en la memoria para optimizar el rendimiento y la escala. (Ayelo & Alberto, 2015)

***Oracle Data Integrator Enterprise Edition y Oracle GoldenGate*:** Mueve y transforma datos estructurados o no estructurados e inmediatamente pasa la información cuando sea necesario y en el formato adecuado para un mejor apoyo en la toma de decisiones. (Ayelo & Alberto, 2015)

***Oracle Business Analytics*:** Permiten realizar análisis en tiempo real y ampliar todo lo que sea posible para la empresa. (Ayelo & Alberto, 2015)

***Oracle Real-Time Decisions y Oracle BPM*:** Proporcionan la acción mediante el apoyo a la toma de decisiones automatizadas, así como las decisiones más complejas para la gestión de procesos de negocio inteligente. (Ayelo & Alberto, 2015)

SAP HANA

Es la aplicación *in-memory* de SAP que consiste en un hardware y un software empaquetado que permite una velocidad de procesamiento nunca visto antes. Permite obtener increíbles tiempos de respuesta al momento de realizar una consulta comparándola con los sistemas de bases de datos tradicionales.

Los servidores de SAP HANA pueden variar en su dimensión dependiendo del requerimiento de la empresa. Las herramientas de explotación de información para HANA brindadas por SAP son: *Business Objects* (BO) *Web Intelligence*, BO *Dashboard Design*, SAP Lumira, SAP *Design Studio* y BO *Analysis.*

La información puede ser explotada por cualquier herramienta que utilice los conectores más utilizados del mercado tales como ODBC y JDBC. Empresas como *MicroStrategy* ya contemplan la conexión directa por lo que se pueden hacer excelentes análisis con una potencia impresionante que permitirá obtener la información a una velocidad imaginada hasta hace poco. (Ayelo & Alberto, 2015)

Amazon

Su portafolio más representativo orientado a *Big Data*, incluye el *Elastic Map Reduce* basado en *Hadoop*, la base de datos *Big Data DynamoDB*, el almacén de datos paralelamente masivo *RedShift*, y todos funcionan bien con *Amazon Web Services.*

Amazon Elastic MapReduce **(Amazon EMR):** Es un servicio web que facilita el procesamiento rápido y rentable de grandes cantidades de datos.

Amazon EMR simplifica el procesamiento de *Big Data* y proporciona un marco de trabajo de *Hadoop* administrado que facilita la distribución y el procesamiento de grandes cantidades de datos entre instancias de Amazon EC2 dinámicamente escalables de manera sencilla, rápida y rentable.

También puede ejecutar en Amazon EMR otros marcos de trabajo distribuidos populares, como Apache *Spark* y Presto, e interactuar con los datos de otros almacenes de datos de AWS, como Amazon S3 y *Amazon DynamoDB.*

Amazon EMR administra con seguridad y fiabilidad sus casos de uso de *Big Data*, incluido el análisis de logs, la indexación web, el almacenamiento de datos, el aprendizaje automático, el análisis financiero, la simulación científica y la bioinformática.(«AWS | *Elastic mapreduce* (EMR) para el procesamiento rápido de datos», s. f.)

***Amazon DynamoDB*:** Es un servicio de datos NoSQL totalmente gestionado y que ofrece un rendimiento rápido y predecible gracias a una perfecta escalabilidad. Este servicio permite a los clientes evitar las cargas administrativas que suponen tener que utilizar y escalar bases de datos distribuidas a AWS, ya que no tiene que preocuparse del aprovisionamiento del hardware, ni tampoco de las tareas de instalación y configuración, replicación, revisiones del software ni de escalar el clúster.(«AWS | Servicio de base de datos gestionada NoSQL (*DynamoDB*)», s. f.)

***Amazon RedShift*:** Es una solución rápida y totalmente gestionada de almacén de datos a escala de *PetaBytes* que permite analizar todos los datos utilizando de forma sencilla y rentable todas las herramientas de inteligencia empresarial que ya disponga.(«AWS | Solución de almacenamiento y análisis de datos en la nube», s. f.)

***Amazon Web Services* (AWS):** Ofrece un conjunto exhaustivo e integral de servicios de informática en la nube para ayudar a gestionar grandes datos gracias a la reducción de costos, escalabilidad para atender la demanda y el aumento de la velocidad de la innovación. (Ayelo & Alberto, 2015)

Microsoft SQL Server

Actualmente SQL Server es la herramienta que propone Microsoft para contener grandes volúmenes de datos con módulos con capacidad analítica, como *Intelligent System Service* (IIS) y *Analytics Platform System* (APS).

Microsoft además propone utilizar en sus Windows Server y sistemas Azure la implementación HD *Insight* que permite recaudar tanto de por ejemplo un sistema *Hadoop* que tiene los datos desestructurados como desde un SQL *Server Parallel Data WareHouse*. (Ayelo & Alberto, 2015)

***Microsoft Azure intelligent System Service* (ISS):** Es el Nuevo servicio de Azure que ayuda a los clientes a adoptar soluciones de *Internet of Things* (IoT) a través de una conexión segura y una gestión y captura de datos ordenada, independientemente de que esa información este generada por maquinas, sensores, dispositivos, etc. y del sistema operativo que se utilice. (Ayelo & Alberto, 2015)

***Analytics PlatForm System* (APS):** Almacena y gestiona los datos tradicionales y también los nuevos tipos de información en un renovado concepto de *Data*

WareHouse. Así, APS combina las mejores funcionalidades de Microsoft SQL y de la tecnología *Hadoop* en un producto de bajo costo. (Ayelo & Alberto, 2015)

Soluciones *Google* para *Big Data*

Es uno de los principales marcadores de tendencias en la revolución de la información en todos los sectores y segmentos.

***BigQuery*:** Es el Servicio Web de *Google* que permite realizar almacenamiento y consulta de datos masivos con billones de filas. Su uso es sencillo y permite a los desarrolladores y analistas de negocios estudiar bases de datos en tiempo real. Realiza consultas del tipo SQL sobre conjuntos de datos que contienen terabytes de información en unos pocos segundos. Conociendo previamente el lenguaje SQL, la programación de consultas es realmente sencilla. Los resultados se pueden almacenar en tablas y también exportar para su análisis externo. Existen suficientes herramientas de terceros que interactúan con *BigQuery* para realizar, carga consultas y visualización de datos. También se pueden utilizar herramientas propias de *Google* como *Prediction* API que utiliza modelos de *Machine Learning* que pueden realizar predicciones en tiempo real. (Ayelo & Alberto, 2015),(Google, s. f.)

Recientemente, *Google* lanzo *Cloud Storage Nearline* que promete tiempos de respuesta de 3 segundos en el acceso a la información con un costo de 1 céntimo de dollar por GB almacenado. Este servicio ofrecería alta seguridad, Backups (para recuperar información ante una falla), redundancia para mayor disponibilidad, sencillez y compatibilidad con la nube de *Google*. (Ayelo & Alberto, 2015),(Google, s. f.)

***TensorFlow*:** Es el sistema de *Machine Learning* de *Google* de segunda generación (el de primera generación se llamaba *DistBelief*) preparado para redes neuronales, inteligencia artificial y *Deep Learning*. El core de *Tensorflow* está escrito en C++, ofrece librerías en Python y C++ y puede correr sobre CPU, GPU, en Linux y Mac, además de en Android e iOS. Además es *Open Source*.(«*TensorFlow* - una biblioteca de software de fuente abierta para la máquina de Inteligencia», s. f.)

VMware

VMware vSpHere Big Data Extensions, el cual permite que *vSpHere* controle las implementaciones de *Hadoop* y hace que lanzar proyectos de *Big Data* se vuelva mucho más sencillo para las empresas. La virtualización de *Hadoop* mediante

vSpHere ofrece nuevos niveles de agilidad para ayudar a implementar, ejecutar y administrar los clústeres de *Hadoop* y, al mismo tiempo, mantener el rendimiento del sistema al mismo nivel que las implementaciones físicas. Mediante la virtualización de *Hadoop* con *vSpHere*, las empresas pueden reasignar los recursos de hardware subutilizados para ejecutar simultáneamente distintos tipos de cargas de trabajo junto a *Hadoop* en un anfitrión físico único. Esto libera los recursos no utilizados y aumenta la utilización del sistema para lograr la eficiencia máxima de hardware. (Ayelo & Alberto, 2015),(«Virtualización de servidor con *VMware vSphere* | *VMware* Colombia», s. f.)

Cloudera

La distribución de Cloudera (CDH) fue la primera en aparecer en el mercado, combinado *Big Data* y *Hadoop*. CDH no solo incluye el núcleo de *Hadoop* (HDFS, *MapReduce*, etc.) sino que también integra diversos proyectos de Apache (*Hbase, Mahout, Pig, Hive*, etc.). CDH es 100 % código abierto, y cuenta con una interfaz gráfica propietaria, Cloudera Manager, para la administración y gestión de los nodos del clúster *Hadoop*. La descarga es totalmente gratuita. No obstante, también cuenta con una versión empresarial, que incluye una interfaz más sofisticada. Cloudera recientemente ha estrechado vínculos con IBM y ORACLE. (Ayelo & Alberto, 2015),(*Cloudera, Hadoop, trademarks, & Here*, s. f.)

Cloudant

Complementa el portafolio de *Big Data* y *Analytics* de IBM más allá de la gestión tradicional de datos, proporcionando base de datos-como-servicio que incluyen una variedad de datos estructurados y no estructurados.(«IBM adquiere Cloudant», 2014),(Romero Albarracín, Vargas López, Rojas Cordero, & Director, 2016)

Los servicios administrados de nube Cloudant logran:

- Almacenar datos de cualquier estructura como documentos auto- descritos de JSON

- Aprovechar un sistema de replicación multi-master y principios de diseño distribuidos avanzados para lograr agrupaciones de bases de datos elásticas que pueden abarcar varios bastidores, los centros de datos o proveedores de la nube.

- Permitir la distribución global de datos así como cargas geo-balanceadas a fin de proporcionar alta disponibilidad y un rendimiento mejorado para aplicaciones que requieren que la información se encuentran cerca de los usuarios.

- Proporcionar búsqueda de texto completo, consulta geo-espacial y temporal avanzada y flexible, así como la indexación en tiempo real.

- Integrarse a través de una interfaz de programación de aplicaciones REST (API).

- Permitir la replicación fácil de datos así como la sincronización para las aplicaciones móviles, con código abierto, librerías de software del dispositivo nativo.

- Ofrecer monitoreo 24x7 y la gestión realizada por sus expertos en Big Data.

HortonWorks

Es otro proveedor de *Hadoop*, es muy conocido por sus alianzas estratégicas con Microsoft, *RackSpace, Red Hat*, TeraData, y otras compañías. Es totalmente *open-source*, con soporte empresarial. Incluye las herramientas que forman el núcleo de *Hadoop*, y por supuesto también incorpora diferentes proyectos *open-source* de Apache. Integra Apache Ambari (no es propietaria) como herramienta de gestión del clúster. (Ayelo & Alberto, 2015)

Splunk

Es un software para buscar monitorizar y analizar datos generados por dispositivos, sistemas e infraestructura IT, a través de una interfaz web. Captura, indexa y correlaciona en tiempo real, almacenándolo todo en un repositorio donde busca para generar gráficos, alertas y paneles fácilmente definibles por el usuario. (Ayelo & Alberto, 2015),(«Inteligencia Operacional, Administración de registros, Administración de aplicaciones, Seguridad y cumplimiento de empresa», s. f.)

MicroStrategy

Su producto Visual *Insight de Business Intelligence*, permite un análisis *in-memory* y mejor conectividad con *Hadoop*, consumiendo de muchas fuentes de datos al mismo tiempo, tales como *DataMarts*, bases de datos, archivos planos entre otros. Además *MicroStrategy* 9.3 ofrece una experiencia de búsqueda tipo *Google* a alta velocidad e introduce un innovador producto de administración, *MicroStrategy*

System Manager, para la automatización de procesos manuales. *MicroStrategy* incluye en su versión 9.3 la capacidad de acceder a la información de negocio almacenada en las fuentes de datos de Apache *Hadoop* y SAP HANA. MicroStrategy 9.3 permite a los usuarios de negocio construir informes, desarrollar análisis y explorar los datos sin necesidad de tener conocimientos tecnológicos. (Ayelo & Alberto, 2015)

MongoDB

Es una base de datos con el perfil NoSQL orientada a documentos, bajo la filosofía de código abierto. La importancia de MongoDB radica en su versatilidad, su potencia y su facilidad de uso, al igual que en su capacidad para manejar tanto grandes como pequeños volúmenes de datos. Es una base de datos que no tiene concepto de tablas, esquemas, SQL, columnas o filas. No cumple con las características de Atomicidad, Consistencia, Aislamiento y Durabilidad. Para almacenar y recuperar los datos hace uso de JSON, pero utiliza BSON, que es una forma binaria de JSON, el cual ocupa menos espacio al almacenar los datos. Otra característica de MongoDB es que realiza consultas dinámicas, es decir, puede realizar consultas sin demasiada planificación. MongoDB se desarrolló en C++.(Data, 2015)

Cassandra

Es una base de datos NoSQL, mayormente desarrollada por Datastax aunque empezó como un proyecto de Facebook, escrita en Java y *open-source* -también es un proyecto Apache-. Entre sus características se encuentra la de ofrecer alta disponibilidad de los datos junto con una gran capacidad para escalar linealmente, además de ser tolerante a fallos -no tiene un punto de fallo único- y compatible con hardware de bajo presupuesto o infraestructuras en la nube.(«El Proyecto Apache Cassandra», s. f.)

Apache *Hadoop*

Apache *Hadoop* es un *framework* de software que soporta aplicaciones distribuidas bajo una licencia libre de la comunidad Apache. Permite el procesamiento de grandes volúmenes de datos de forma distribuida a través de clústeres usando modelos sencillos de programación. Está siendo construido y usado por una comunidad global de contribuyentes, mediante el lenguaje de programación Java. Está diseñado para escalar desde un servidor sencillo hasta miles de nodos los cuales pueden ser heterogéneos.(«Welcome to Apache™ Hadoop®!», s. f.)

Apache *Spark*: Es un *framework* de computación en paralelo que genera velocidades hasta 100 veces mayores que las desarrolladas por *Hadoop MapReduce* en memoria, o 10 veces más rápido en el disco, relacionados con el procesamiento de datos a gran escala. Es considerado la evolución de *Hadoop*.

Spark puede ejecutar análisis más rápidos que los despliegues de *Hadoop* existentes, puede coexistir con las instalaciones existentes de *Hadoop* y añadir nuevas funcionalidades. *Spark* se integra perfectamente con *Hadoop* usando YARN de *Hadoop* 2.0. Además puede funcionar con muchos otros productos de Big Data como: CassandraDB, *Google Big Query*, almacenamiento de Amazon S3, *Elastic Search*, etc. Es capaz de usar fuentes de datos existentes (SQL, HIVE, CassandraDB, MongoDB, JDBC, etc.), se puede usar para gestionar las fuentes internas de datos (RDDs - *DataFrames*) como tablas estructuradas.

Spark fue desarrollado en un nuevo lenguaje funcional y orientado a objetos *Scala*. Gracias a *Scala* se puede programar de manera muy concisa y fluida soluciones que antes requerían cientos de líneas. Además se puede programar en Python, R e incluso en Java.

Apache *Spark* es una herramienta útil y eficiente para tareas de procesamiento de *streaming, machine learning* (MLlib), cálculo de grafos (GraphX), integración con lenguaje R y análisis interactivos. Está en constante desarrollo y se actualiza frecuentemente. Además, su documentación es muy completa y la comunidad cada vez se hace más grande. («Apache *Spark*™ - *Lightning-Fast Cluster Computing*», s. f.)

Apache *Mahout*: Es un librería *open source* de Apache Software *Foundation* (ASF) cuyo objetivo principal es la producción de implantaciones libres de algoritmos escalables de *machine learning* usando el paradigma *MapReduce. Mahout* contiene implementaciones para filtrado colaborativo, *clustering* y clasificación. *Mahout* también proporciona librerías de Java para operaciones matemáticas comunes (centradas en el álgebra lineal y estadísticas) y colecciones de primitivas Java.(«Apache *Mahout*: *Scalable machine learning and data mining*», s. f.)

Apache *Zeppellin*: Es una implementación del concepto de web notebook, centrado en la analítica de datos interactivo mediante lenguajes y tecnologías como *Shell, Spark, SparkSQL, Hive, Elasticsearch*, R, entre otros. Es una “interfaz de usuario basada en navegador que ofrece una capacidad de estilo

portátil para los analistas y científicos de datos para explorar sus datos interactivamente y realizar análisis de datos sofisticados."(«*Hadoop Summit*: la seguridad un reto de *Big Data* según *Hortonworks*», s. f.), posibilitando la obtención de datos de múltiples fuentes y la utilización de diferentes tecnologías y de lenguajes de programación.(«Zeppelin», s. f.)

MapR Technologies

MapR es una distribución para *Hadoop* que pone a disposición de los usuarios tres versiones diferentes de su producto. De ésta manera consiguen ofrecer un producto adaptable a las necesidades del usuario. Las tres versiones de MapR son: MapR M3, MapR M5 y MapR M7(Guevara & Antonio, 2015). La versión M5 contiene toda la versión M3 más algunas funcionalidades extras. Mientras que la versión M7 engloba todas las funcionalidades de la M5 y añade de nuevas. La versión M3 es básicamente un paquete con todo el *framework Hadoop* y el ecosistema, preparado y probado para facilitar al usuario la instalación de un clúster

Esta distribución posee una gama de soluciones con respecto al desarrollo sobre un ecosistema *Hadoop*. MapR ofrece soluciones para trabajar con bases de datos SQL sobre *Hadoop*, como también ofrece soluciones NoSQL.(«MapR: Plataforma de datos convergente», s. f.)

Pentaho Business Intelligence

La suite de *Pentaho BI* brinda un conjunto de potentes herramientas capaces de cubrir todo el desarrollo de un proyecto de BI. Desde la integración de datos, haciendo uso de la herramienta *Pentaho Data Integration* pasando por la minería de datos con el software *Weka* incluido en el proyecto *Pentaho*, hasta la implementación y desarrollo con la herramienta Bi Server y los múltiples componentes de los que dispone. Todo esto libre de licencias y con una comunidad de usuarios detrás, que asegura el soporte y la actualización a corto-mediano plazo.(Guevara & Antonio, 2015),*(«Pentaho | Data Integration and Business Analytics Platform for Big Data Deployments*», s. f.)

Los módulos incluidos por la suite son: («*Pentaho | Data Integration and Business Analytics Platform for Big Data Deployments*», s. f.),(«*Desarrollo Pentaho - Intryo*», s. f.)

***Pentaho Data Integration* (PDI):** Transforma e integra datos entre sistemas de información existentes y los *Datamarts* que conformaran el sistema BI.

***Pentaho Reporting* (PRD):** Obtiene y muestra informes de los indicadores de la organización.

***Pentaho Analysis*:** Permite consultar, explorar y analizar la información de empresa de manera interactiva, pudiendo seleccionar diferentes perspectivas de dicha información en base a criterios predefinidos.

***Pentaho Data Mining*:** Facilita el descubrimiento de patrones de comportamiento e indicadores ocultos en la información de la compañía.

***Sparkl*:** Potente herramienta para la creación de *plugins* personalizados.

***Big Data*:** Conjunto de herramientas necesarias para poder analizar grandes volúmenes de datos con el objeto de poder identificar patrones recurrentes en estos datos. PDI incorpora los conectores necesarios para los principales distribuidores de bases de datos analíticas.

SAP *Business Intelligence*

En este caso no se trata de una herramienta, si no de varias. SAP ofrece múltiples soluciones(«MapR: Plataforma de datos convergente», s. f.), cada una adaptada a unas características determinadas, abarcando una gran variedad de fuentes de datos e integración con software de terceros. Cabe destacar que las múltiples necesidades de software de una empresa pueden ser cubiertas mediante el software SAP, que ofrece soluciones para múltiples sistemas de información:

SAP *Business Information Warehouse* (SAP BW): Permite analizar datos operativos de aplicaciones SAP, de otras aplicaciones de negocios y de fuentes externas como bases de datos, servicios on-line e Internet. Asimismo, SAP BW está pre configurado por áreas y procesos, permitiendo examinar las relaciones existentes entre todas las áreas de la organización.

SAP *Knowledge Management* (SAP KM): Esta herramienta proporciona al usuario la capacidad de buscar, acceder, crear y difundir la información más allá de los límites de los repositorios donde se encuentra la información. Permitiendo que la información se integre dentro de las aplicaciones y procesos de negocio, proporcionando un valor mayor al contenido de la información.

SAP *Strategic Enterprise Management* (SAP SEM): permite cambiar dinámicamente las estrategias.

Proyecto R

Se trata de un proyecto de software libre, resultado de la implementación GNU del lenguaje S. Es un lenguaje y entorno de programación para análisis estadístico y gráfico, el cual brinda una excelente estructura para relacionar con almacenes de *Big Data* para plasmar gráficamente los análisis realizados de un conjunto de datos.

R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos. Entre otras características dispone de: Almacenamiento y manipulación efectiva de datos, operadores para cálculo sobre variables indexadas (*Arrays*), una amplia e integrada colección de herramientas para análisis de datos, posibilidades gráficas para análisis de datos y lenguaje de programación bien desarrollado, simple y efectivo. R es en gran parte un vehículo para el desarrollo de nuevos métodos de análisis interactivo de datos.(Romero Albarracín et al., 2016)

QlikView

Herramienta de *Business Intelligence* de gran flexibilidad, con posibilidad de obtener soluciones adaptadas a áreas determinadas. A destacar su compatibilidad con gran variedad de fuentes de información con las que puede trabajar, incluyendo hojas de cálculo de Excel. Permite la creación de interfaces intuitivas con gráficos a través de los que se puede navegar y acceder a diferente información mediante filtros automáticos. Presenta también propuestas para la integración del entorno móvil y la creación de apps para el mismo.

Su tecnología basada en la lógica asociativa posibilita realizar cálculos en tiempo real, que permiten a los profesionales de los negocios tener un conocimiento profundo de su empresa gracias a exploraciones intuitivas de datos. Puede conectarse a cualquier origen de datos (Oracle, AS/400, Sybase, Db2, Informix, xml, mdb, xls, csv), directamente o a través de conexiones ODBC y OLEDB.

Estos datos se almacenan en una BBDD propietaria que se crea automáticamente y con tecnología AQL (*Associative Query Language*). *QlikView* construye un modelo analítico dinámico que no limita las posibilidades en la explotación de la información.(«*Guided Analytics* | *Business Intelligence Software* | *QlikView*», s. f.),(Contel Rico, 2011)

Aspectos importantes a tener en cuenta:(«Comparativa y diferencias entre las herramientas de *Business Intelligence Pentaho y Qlikview*», s. f.)

- Utiliza la lógica asociativa (AQL), técnica que realiza los análisis y cálculos en memoria con lo cual obtiene tiempos de respuesta excelentes.

- Su costo es sensiblemente inferior respecto a Business *Objects, Cognos, Microstrategy*, etc.

- Sus Cuadros de Mando son elegantes y sencillos de usar. Carece de metadatos centralizados

- Al usar lógica asociativa, no dispone de una suite ETL.

- El tiempo de implementación suele ser inferior a 3 meses, por lo tanto, el costo de consultoría es menor.

- Curva de aprendizaje inferior a 1 semana.

- Accesibles desde iPad, Android, etc.

- Dispone de interfaces gráficas y *Wizards* muy intuitivos.

- Costo por licencia

- El soporte está incluido y es brindado por el fabricante.

- Entornos de escritorio, web y móvil

Pivotal

Es una compañía fundada por EMC, *Greenplum* y en colaboración con *VMWare*. Está centrada en soluciones *Big Data* -con Pivotal HD- y en soluciones *cloud* -con Pivotal CF.

Pivotal HD *Enterprise* es la solución construida sobre *Hadoop* 2.0 cuenta con HDFS como almacenamiento principal y con MRv2 más las principales herramientas de análisis de datos: *Pig, Hive, Mahout*… Las principales adiciones son la herramienta de administración y monitorización *Command Center*, el paquete HAWQ, que incluye una herramienta de consulta de datos MPP y una librería de funciones para el análisis de datos; y el paquete *GemFire*.

Adicionalmente se incluyen otras herramientas como el *Data Loader*, para cargar datos a HDFS; o Spring, para realizar *workflows*.

Pivotal HD cuenta con una versión gratuita -*Community*- pero está limitada sólo a una distribución *Hadoop* 2.0 y a algunas herramientas y utilidades aportadas por Pivotal como el *Data Loader* y *Command Center*. Además de una limitación para los clústeres de 50 nodos. Esta versión sin embargo no incluye HAWQ, *GemFire* XD ni el paquete de funciones para análisis de datos.

Pivotal CF es una plataforma PaaS (plataforma como servicio) de código abierto, ofrece a sus usuarios una selección de nubes, el desarrollo y la infraestructura de servicios de aplicaciones que les permitan crear, probar y desarrollar aplicaciones. Se distribuye bajo la licencia Apache.(Software, 2015)

SPSS

Es un programa estadístico informático muy usado de IBM en las ciencias sociales y las empresas de investigación de mercado. Facilita el análisis estadístico y presentación de informes, aborda todo el proceso analítico: planificación, recopilación de datos, análisis, informes y despliegue. También permite la construcción de modelos predictivos y minería de datos. Analiza grandes volúmenes de datos para obtener información predictiva y construir estrategias comerciales eficaces. Tiene tres modelos de implementación *OnPremise*, *Cloud* o Hibrido.(«IBM SPSS - IBM *Analytics*», 2016)

Recientemente ha sido desarrollado un paquete libre llamado PSPP, con una interfaz llamada PSPPire que ha sido compilada para diversos sistemas operativos como Linux, además de versiones para Windows y OS X. Este último paquete pretende ser un clon de código abierto que emule todas las posibilidades del SPSS.(«PSPP», s. f.)

Tableau

Es una herramienta de Inteligencia de Negocios que permite visualizar grandes volúmenes de información en forma rápida, flexible y amigable. Es destacada por su facilidad de uso, potencialidad para generar visualizaciones y capacidad de manejo de grandes volúmenes de Datos.

A diferencia de las herramientas tradicionales de Inteligencia de Negocios (BI) desarrolladas pensando en el usuario técnico del área de sistemas, Tableau está orientado a que personas de todos los ámbitos puedan manejar información

fácilmente y presentarla en forma atractiva. Encontrando un poderoso aliado analítico.

Tableau es rápido, ofrece flexibilidad en la manera en que se conecta a fuentes de datos. R es un reconocido lenguaje y entorno de programación para análisis estadístico y gráfico. Tableau *Desktop* ahora puede conectarse a este entorno mediante campos calculados y aprovechar las funciones, bibliotecas, paquetes y modelos disponibles. Ofrece tres versiones *Desktop*, *Server* y *Cloud*.(«Tableau Software», s. f.)

MLlib

Es la librería escalable de *machine learning* de *Spark*. Proporciona múltiples tipos básicos de algoritmos y utilidades como: clasificación, regresión, *clustering*, filtrado colaborativo, reducción de dimensionalidad y optimización de primitivas. Además aporta varias estadísticas básicas como: Resumen de estadísticas por columnas, Correlaciones, Muestreo estratificado, Pruebas de hipótesis y Generación de datos aleatorios. («MLlib | Spark Apache», s. f.)

Weka

Weka es un acrónimo de *Waikato Environment for Knowledge Analysis*, es un entorno para experimentación de análisis de datos, escrito en java, que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario como: pre procesamiento de datos, agrupamiento, clasificación, regresión, visualización y características de selección. Sus técnicas se basan en la hipótesis de que los datos están disponibles en un único archivo plano o relación, donde cada punto marcado es etiquetado por un número fijo de atributos.

WEKA proporciona acceso a bases de datos SQL utilizando conectividad de bases de datos Java y puede procesar el resultado devuelto como una consulta de base de datos. Su interfaz de usuario principal es el Explorer, pero la misma funcionalidad puede ser accedida desde la línea de comandos o a través de la interfaz de flujo de conocimientos basada en componentes.(«Weka 3 - *Data Mining with Open Source Machine Learning Software in Java*», s. f.)

Microsoft Power BI

Es un conjunto de aplicaciones de análisis de negocios que permite analizar datos y compartir información. Puede acceder a sus datos e informes desde cualquier

lugar con las aplicaciones móviles de *Power BI Mobile*, que se actualizan automáticamente con los cambios que se realizan en los datos. *Power BI* puede unificar todos los datos de su organización, ya sea en la nube o localmente, puede conectar bases de datos SQL Server, modelos de *Analysis Services* y muchos otros orígenes de datos a los mismos paneles en *Power BI.*

Hay dos versiones que el usuario puede adquirir una de ellas es la versión *POWER BI Desktop* que es gratis pero con ciertas restricciones de uso y la otra versión es la versión Cloud cuyo licenciamiento tiene costos muy competitivos por mes cada licencia.(«*Power BI* | Herramientas de BI para la visualización de datos interactivos», s. f.)

Uno de los aspectos importantes a tener en cuenta es el tiempo de implementación estimado en una semana(«Un nuevo punto de Vista sobre Inteligencia de Negocios – Microsoft Power BI», s. f.)

Tibco SpotFire

Es una plataforma de análisis empresarial orientada a usuarios técnicos y de negocio, que integra potentes funcionalidades para el análisis de autoservicio, visuales, interactivos e intuitivos, que proveen excelentes criterios de estudio para anticipar tendencias y patrones emergentes, permitiendo unificar la creación y entrega de informes desde el punto de vista colaborativo contextual e interoperable. Las siguientes son las herramientas que integra:

- ***Analyst*:** Realiza informes analíticos integrales. Crea indicadores claves de desempeño y métricas para usuarios móviles.
- ***Business Autor*:** Cliente basado en web orientado a usuarios avanzados para la creación de reportes creativos predictivos.
- ***Consumer*:** Interfaz web colaborativa para la navegación por flujos de trabajo de análisis integrados y configurados en plataformas sociales.
- ***Statiscal service*:** cuantifica y minimiza la incertidumbre en modelos estadísticos y predictivos, basados en TERR, R, S+, SAS y MATLAB.
- ***Data connectivity*:** Permite extraer datos de cualquier fuente de información y sistemas transaccionales. No requiere soporte TI

Es un software de análisis que permite extraer información valiosa rápidamente para tomar decisiones más acertadas.(Morales & Carolina, 2015),(«*Data Visualization & Analytics Software - TIBCO Spotfire*», s. f.)

ANEXO D. ESCALA DE VALORACIÓN PARA SOLUCIONES

La escala utilizada para la valoración de cada uno de los aspectos evaluados fue la siguiente: 1 (Deficiente), 2 (Aceptable) y 3 (Excelente). A continuación se explica el significado de cada parámetro:

- **Costo de implementación:** El valor de la adquisición de la solución. No se detalla un precio exacto sino el tipo de gasto que implica la compra de la infraestructura. Los valores para este parámetro son:
  1: El valor es alto muy poco flexible.
  2: Tiene un valor alto pero es flexible a las necesidades del proyecto.
  3: El valor es totalmente flexible a las necesidades del proyecto.

- **Costo de mantenimiento:** Los gastos asociados a mantener la solución. No se detalla un valor exacto sino el tipo de gasto que implica soportar la infraestructura. Los valores para este parámetro son:
  1: El valor es alto muy poco flexible.
  2: Tiene un valor alto pero es flexible a las necesidades del proyecto.
  3: El valor es totalmente flexible a las necesidades del proyecto.

- **Facilidad de Instalación y configuración:** Nivel de conocimientos necesarios para la instalación y configuración de las soluciones. Los valores para este parámetro son:
  1: Requiere nivel de conocimientos alto en hardware y software.
  2: Requiere nivel de conocimientos medio en instalación de software.
  3: Transparente para el usuario. No requiere conocimientos.

- **Escalabilidad:** Las facilidades de cada infraestructura para cambiar el tamaño de los clústeres. Los valores para este parámetro son:
  1: Permite la escalabilidad pero está restringida a un cierto tipo de hardware.
  2: Permite la escalabilidad sin importar el hardware, requiere conocimiento.
  3: Permite la escalabilidad es sencillo, no implica trabajar con el hardware.

- **Flexibilidad:** La posibilidad de configurar las opciones que ofrecen a nivel de software. Los valores para este parámetro son:
  1: Solo permite trabajar con el software que incluye.
  2: Permite trabajar con otro software. Tiene restricciones de Compatibilidad.
  3: Es totalmente abierto a la configuración de software.

- **Rendimiento:** Desempeño de las diferentes arquitecturas según los tipos de solución. Los valores para este parámetro son:
  1: Bajo desempeño.
  2: Medio desempeño.
  3: Alto desempeño.

- **Experimentación:** Posibilidad de experimentar y hacer pruebas en la solución tanto en hardware como en software, antes y después de la adquisición. Los valores para este parámetro son:
  1: Permite experimentación restringida y solo en el entorno definitivo.
  2: Permite experimentación pero en un entorno de producción.
  3: Permite experimentación e incluso es posible un entorno de prueba.

ANEXO E. ESCALA DE VALORACIÓN PARA DISTRIBUCIONES

La evaluación realizada se hace basada en una comparación teórica de sus características y funcionalidades. Los parámetros seleccionados han sido calificados como: 1 (Bajo), 2 (Medio) y 3 (Alto), descritos a continuación:

- **Plataforma:** Indica las plataformas sobre las cuales puede ser configuradas la distribución. Los valores para este parámetro son:
  1: Corre sobre una sola plataforma, básicamente Linux.
  2: No aplica
  3: Corre sobre Linux y Windows Server

- **Presencia en el mercado:** La aceptación y uso que tiene la distribución en el mercado entre las empresas. Es un factor difícil de valorar pero importante. en las valoraciones. Los valores para este parámetro son:
  1: Tiene poca presencia ya que es recién llegada a la industria *Big Data*.
  2: Utilizada por empresas grandes, no cuenta con comunidad de respaldo.
  3: Utilizada por empresas grandes, cuenta con comunidad de respaldo

- **Versión gratuita:** para efectos de pruebas de concepto es conveniente siempre tener la opción de probar sin pensar en costos. Los valores para este parámetro son:
  1: Incluyen una distribución *Hadoop* y algo más.
  2: Incluyen gran parte de las herramientas pero con limitaciones.
  3: Incluye gran parte de las herramientas.

- **Componentes *Hadoop*:** Se refiere a la cantidad de herramientas *open source* que cada distribución ofrece. Los valores para este parámetro son:
  1: Incluyen mínima cantidad de herramientas *open source*.
  2: Incluyen mediana cantidad de herramientas *open source*.
  3: Incluyen gran cantidad de herramientas *open source*.

- **Administración:** Hace referencia a las herramientas de administración que ofrece cada distribución. Los valores para este parámetro son:
  1: Incluyen las mínimas herramientas de administración.
  2: Incluyen un nivel intermedio de herramientas de administración.
  3: Incluyen gran variedad de herramientas de administración.

- **Productividad y desarrollo:** Se indican las facilidades que ofrece una distribución a la hora de trabajar con ella. Los valores para este parámetro son:
  1: Ninguna facilidad para desarrollar análisis de los datos almacenados.
  2: Facilita el desarrollo de aplicaciones.
  3: Facilita en gran medida el desarrollo de aplicaciones.

- **Rendimiento:** Evalúa las herramientas adicionadas a la base de *Hadoop* para mejorar su rendimiento. Los valores para este parámetro son:
  1: Ninguna herramienta adicional que mejore el rendimiento del sistema.
  2: Incluye mejoras en el rendimiento de herramientas existentes de *Hadoop*.
  3: Incluye mejoras en el rendimiento. Además de nuevas herramientas.

- **Tolerancia a fallos:** Mide cada distribución con las características disponibles para solucionar los problemas de tolerancia a fallos. Los valores para este parámetro son:
  1: Sin características que aumenten la tolerancia a fallos.
  2: Agrega mejoras a *Hadoop* sobre HDFS o *MapReduce*.
  3: Agrega funcionalidades a *Hadoop* para aumentar la tolerancia a fallos.

ANEXO F. ESCALA DE VALORACIÓN PARA HERRAMIENTAS BUSSINESS INTELLIGENCE

Los parámetros seleccionados han sido calificados como: 1 (Bajo), 2 (Medio) y 3 (Alto), descritos a continuación:

- **Costo licencia:** Indica el valor de una licencia *Cloud* por usuario. Los valores para este parámetro son:
  1: Precio Alto de 1 licencia *Cloud* por mes (Mayor a 50 USD)
  2: Precio Medio de 1 licencia *Cloud* por mes (Entre 9 y 50 USD)
  3: Precio Bajo de 1 licencia *Cloud* por mes (Menos de 9 USD)

- **Comunidad de respaldo:** Hace referencia al tamaño de la comunidad de usuarios que utiliza la herramienta. Los valores para este parámetro son:
  1: Comunidad de usuarios Baja (menos de 100.000 usuarios)
  2: Comunidad de usuarios Media (Entre 100.000 y 499.999 usuarios)
  3: Comunidad de usuarios Alta (MAS de 500.000 usuarios)

- **Curva de Aprendizaje:** Tiene que ver con el tiempo necesario para conocer la herramienta y obtener los primeros resultados:
  1: Curva de aprendizaje Baja (más de tres semanas)
  2: Curva de aprendizaje Media (entre una y dos semanas)
  3: Curva de aprendizaje Alta (menos de una semana)

- **Tiempos de Implementación:** Indica el tiempo necesario para implementar una solución:
  1: Tiempo de implementación Baja (más de tres meses)
  2: Tiempo de implementación Media (entre dos y tres meses)
  3: Tiempo de implementación Alta (menos de 2 meses)

- **Socios de Negocio:** Tiene que ver con la cantidad de empresas distribuidoras de la herramienta en Colombia:
  1: Socios de Negocio Baja (entre 0 y 1 socios)
  2: Socios de Negocio Media (entre 2 y 3 socios)
  3: Socios de Negocio Alta (más de 3 socios)

- **Transformación de Datos:** Hace referencia a si la herramienta incluye procesos de transformación de datos:
  1: No tiene procesos de transformación de datos
  2: Dispone de procesos de Transformación de datos
  3: Los procesos de transformación de datos son una fortaleza.

- **Conectores a Fuentes de Datos:** Tiene que ver con los conectores disponibles a fuentes de datos:
  1: Limitada conectividad
  2: Dispone de conectores nativos y conectores ODBC
  3: Alta conectividad nativa y mediante conectores ODBC

- **Componentes *Hadoop*:** Se refiere a la facilidad de integración de la herramienta con componentes del ecosistema de *Hadoop*
  1: No dispone de integración con componentes *Hadoop*
  2: Dispone de algún tipo de integración con componentes *Hadoop*
  3: Alta integración con componentes *Hadoop*