Universidad Autónoma del Estado de México

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

CENTRO UNIVERSITARIO UAEM ATLACOMULCO

ANTOLOGÍA GESTIÓN Y ANÁLISIS DE BIG DATA LIAD6 2022-A

UNIDAD DE APRENDIZAJE:

GESTIÓN Y ANÁLISIS DE BIG DATA

PROGRAMA EDUCATIVO

LICENCIATURA EN INFORMÁTICA ADMINISTRATIVA

PERIODO

2022 A

COMPILADORES:

MAN. CARLOS ALBERTO BALTAZAR VILCHIS

M.I. ELIZABETH EVANGELISTA NAVA

DRA. EN A. YENIT MARTÍNEZ GARDUÑO

DR. EN C.F. ALBERTO GARDUÑO

FECHA DE ELABORACIÓN:

JULIO 2022

Centro Universitario UAEM Atlacomulco

Antología Gestión y Análisis de Big Data LIAD6 2022-A

Obra compilada por académicos del Centro Universitario UAEM Atlacomulco

MAN. Carlos Alberto Baltazar Vilchis

M.I Elizabeth Evangelista Nava

Dra. en A. Yenit. Martinez Garduño

Dr. en C.F. Alberto Garduño Martínez

Periodo de compilación: Febrero – Julio 2022

Hecho en México

Índice

I. Presentación

Las tecnologías de información han evolucionado hasta el grado de permitir que las máquinas realicen procesos simulando sistemas de información administrativos, se pretende que los alumnos aprendan a través de casos prácticos cuales son las herramientas y técnicas computacionales-administrativas avanzadas para el control de grandes cantidades de datos.

La Unidad de Aprendizaje “Gestión y Análisis de Big Data” tiene la finalidad de desmenuzar macrodatos, datos masivos, inteligencia de datos o datos a gran escala como un concepto que hace referencia a conjuntos de datos tan grandes.

Lo que hace que Big Data sea tan útil para muchas empresas es que proporciona respuestas a muchas preguntas que las empresas ni siquiera saben que se están haciendo. En otras palabras, proporciona un punto de referencia. Con tanta información, los datos se pueden moldear o examinar de cualquier forma que la empresa considere adecuada. De esta forma, las organizaciones pueden identificar los problemas de una forma más comprensible.

Por lo anterior la Unidad de Aprendizaje “Gestión y Análisis de Big Data” pretende diseñar y justificar el aseguramiento de la integridad y confidencialidad de la información por medio de herramientas como ingeniería de software, ciberseguridad, sistemas distribuidos para desarrollar proyectos que incrementen la seguridad y productividad de los sistemas de información de una organización.

Aquí radica la importancia que los sistemas de información tienen para los Administradores, Contadores e Informáticos.

II. Mapa Curricular Licenciatura en Informática Administrativa, 2018

La Unidad de Aprendizaje (UA) Gestión y Análisis de BIG DATA, se imparte en el plan de estudios de la Licenciatura en Informática Administrativa, reestructuración f18 en el octavo periodo, es un curso tipo taller, del núcleo de formación integral.

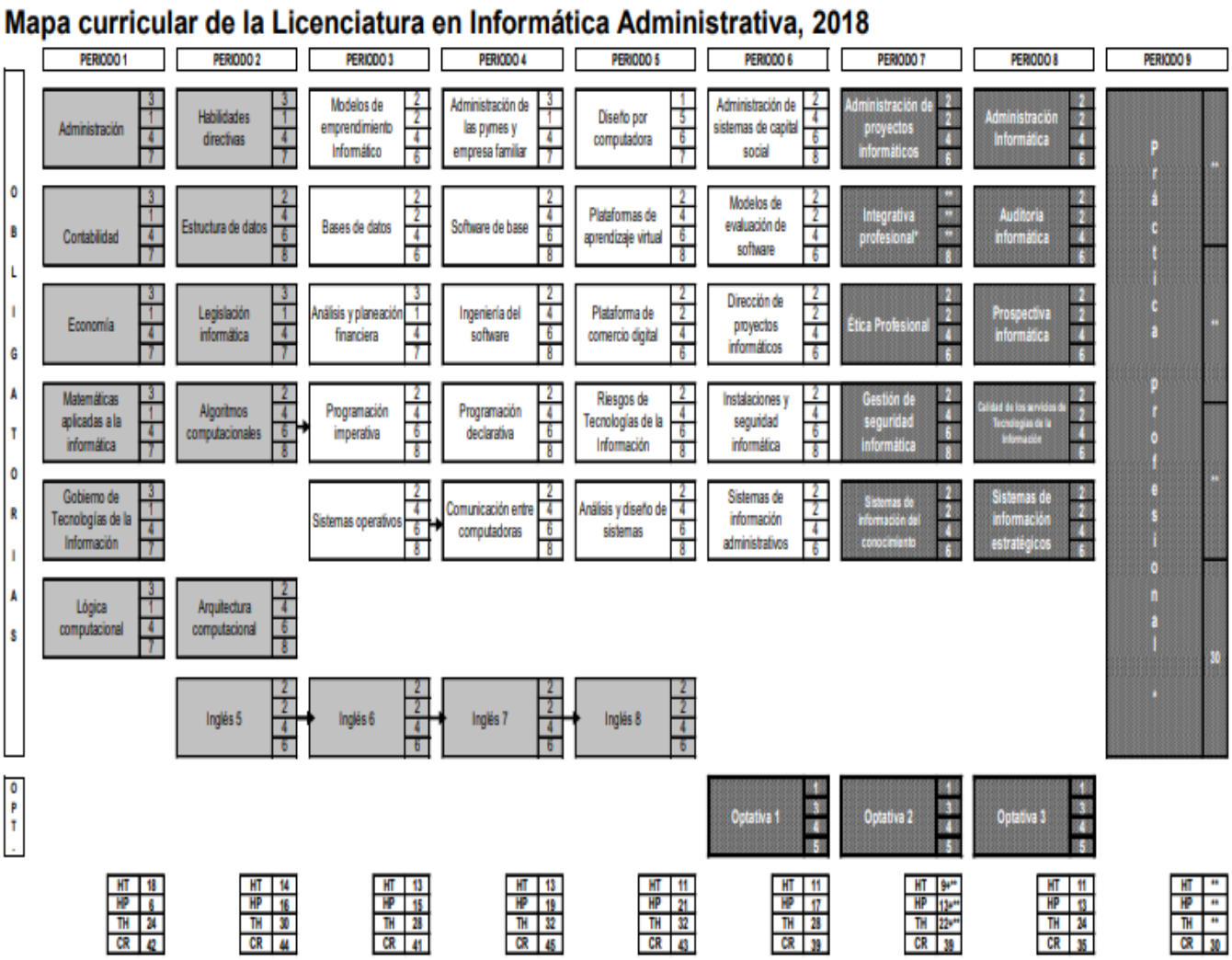


Figura 1. Mapa Curricular UA obligatorias (2018)

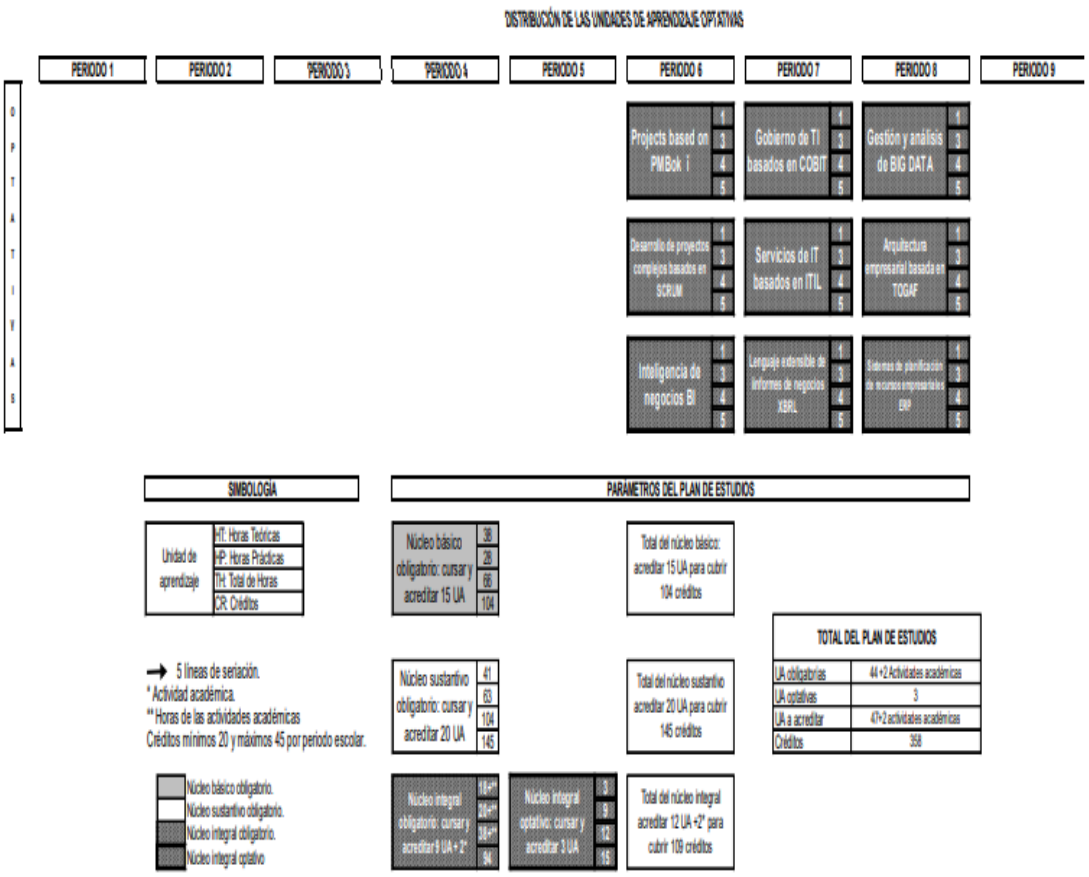


Figura 2. Mapa curricular UA optativas del Plan de estudios LIA (2018)

III. Material referencial al módulo de la unidad de aprendizaje

Tabla 1. Módulo de aprendizaje

UNIDAD DE APRENDIZAJE	TEMA
UNIDAD 1. Conceptualización de BIG DATA	1.1 Antecedentes de BIG DATA 1.2 Conceptos de BIG DATA. 1.3 Análisis y gestión de procesos orientados a la experiencia del cliente 1.4 Estrategias CRM como base del negocio. 1.5 Big Data y Data Managemment. 1.6 Integrar y Conocer las fases de big data y aspectos Legales.
UNIDAD 2. Análisis y Estadística	2.1 Técnicas del tratamiento de datos 2.2 Técnicas analíticas en Big Data. 2.3 Data science. 2.4 Datamining. 2.5 Big Data Analytics.
UNIDAD 3. Desarrollo del Big data.	3.1 Diseños de investigación. 3.2 Aplicaciones de Negocio Big Data. 3.3 Métricas y objetivos específicos del Big Data en el negocio. 3.4 Los tableros de control y reportes Dashboard. 3.5 Business Inteligencie y Big Data como estrategia del negocio. 3.6 Procesado de datos. 3.7 Arquitectura Big Data. 3.8 Learning Big Data. 3.8.1 Datamining. 3.8.2 Socialmining.
UNIDAD 4. Aplicaciones, plataformas y tecnologías en Big data.	4.1. Sistemas operativos, servidores y redes. 4.2 Lenguajes de programación Python y R. 4.3 Almacenamiento y procesamiento de la información en Big Data. 4.3.1 Las 7Vs. 4.3.2 Tipos de Datos para Almacenar (Multivariedad de datos en la BigData). 4.3.3Sistemas de almacenamiento distribuido. 4.3.4 Introducción Hadoop. Arquitectura Hadoop. 4.4 HDFS. Bases de datos SQL y noSQL. 4.5 Aplicación práctica.

UNIDAD I. Conceptualización de BIG DATA

OBJETIVO: Comprender los conceptos del paradigma de la programación declarativa, los elementos de arquitectura de aplicaciones web y los fundamentos para la administración de su desarrollo.

1.1 **Antecedentes del BIG DATA.**

Uno de los primeros términos clave que nos encontramos en el campo del análisis de datos de negocio es el de Business Intelligence, cuya primera referencia se remonta a 1958 a cargo de Hans Peter Luhn (Citado por Niño *et al*, 2015) investigador de IBM, aunque en dicha referencia el término aún estaba alejado de la evolución que sufrió posteriormente con la progresiva informatización de los procesos de negocio. Tras el desarrollo en dicho campo en los años siguientes, es en los años 80 cuando se consolida la idea de Business Intelligence (principalmente con la propuesta de Howard Dresner (Martens, 2006 citado por Niño *et al*, 2015) para referirse a un conjunto de sistemas software para el apoyo a la toma de decisiones de negocio, basados en la recogida de análisis de hechos o datos. Estos sistemas se enfocan en un análisis de tipo descriptivo, consultando datos históricos de manera agregada y cruzando indicadores para obtener una mejor visión de lo que ha pasado y está pasando en la organización. El enfoque de Business Intelligence deja al margen, por tanto, un análisis de tipo predictivo que busca la extracción de conocimiento de los datos en forma de patrones, tendencias o modelos que permitan una cierta certeza sobre el resultado de potenciales acciones futuras.

Para denominar este tipo de análisis, a finales de los 80 surge la expresión Data Mining (minería de datos). El origen del término proviene de la analogía con las técnicas de minería en las que se extrae un material valioso (en este caso, conocimiento) a partir de yacimientos (bancos de datos). Junto al término Data Mining, como probablemente el más conocido y utilizado para referirse a este tipo de análisis de entre un conjunto de expresiones similares (Han, J., & Kamber, M., 2006 citado por Niño *et al*, 2015) en la misma época empieza también a utilizarse la expresión Knowledge Discovery in Databases (KDD). De hecho, el primer seminario académico sobre esta materia se organiza en 1989 (Piatetsky-Shapiro, G. (1991), que en 1995 deriva en la First International Conference on Knowledge Discovery and Data Mining. El desarrollo de proyectos de Data Mining para la búsqueda y explotación de patrones en bancos de datos, empleando técnicas de Machine Learning (aprendizaje automático) para la construcción de modelos predictivos, comienza a extenderse en contextos de negocio durante la década de 1990, destacando su aplicación en el sector bancario y de seguros, donde se busca emplear los resultados de este tipo de análisis para facilitar procesos de toma de decisiones ligadas a productos de dichos sectores (por ejemplo, la detección de fraudes a compañías aseguradoras, o para la concesión o denegación de créditos). Este tipo de aplicaciones de Data Mining nos permite explicar el origen del concepto de Data Science (ciencia de datos). Dicho término surgió a principios de milenio

para denominar una propuesta de revisión de las áreas técnicas en torno a la Estadística, y así adecuarse mejor a las prácticas de análisis de datos que venían desarrollándose en la época, principalmente con el desarrollo del Data Mining y su aplicación en diferentes contextos de negocio, y con la progresiva informatización de la recolección y análisis de datos. A partir de dicha propuesta, el concepto de Data Science ha ido desarrollándose como la integración de principios de las diferentes disciplinas (estadística y matemáticas, informática y computación, fundamentos del área específica de aplicación) que sustentan la práctica moderna del análisis de datos y guían la extracción de conocimiento de los mismos. De esta manera, podemos entender el Data Mining como la extracción en sí de ese conocimiento a través de herramientas y técnicas que incorporan los principios de la ciencia de datos.

1.2 **Conceptos de BIG DATA.**

El Instituto Global de McKinsey (Joyanes, 2013, p.26) hace referencia al término **Big Data** como a los conjuntos de datos cuyo tamaño está más allá de las capacidades de las herramientas típicas de software de bases de datos para capturar, almacenar, gestionar y analizan.

“Podemos denominar **Big Data** como el análisis y gestión de grandes volúmenes de datos los cuales no pueden ser tratados de la manera convencional, y los cuales deben cumplir con la ley de las 4V´s del Big Data.” (Sánchez, D.F,2018, p.9), “La **gestión de información** comprende las actividades relacionadas con la obtención de la información adecuada, a un precio adecuado, en el tiempo y lugar adecuado, para tomar la decisión adecuada.” (Aja, L, 2002).

Joyanes (2013) define, los **Datos Estructurados** como “Datos con formato o esquema fijo que poseen campos fijos.”; los **Datos Semiestructurados** según sean “Datos que no tienen formatos fijos, pero contienen etiquetas y otros marcadores que permiten separar los elementos dato.”; y los **Datos No Estructurados** como “datos sin tipos predefinidos. Se almacenan como “documentos” u “objetos” sin estructura uniforme, y se tiene poco o ningún control sobre ellos”.

Las tres “uves” de *Data Bases* funcionan como atributos que caracterizan a los *macrodatos*, el **Volumen** implica el tamaño y tipo de cantidades de datos y metadatos, la **Velocidad** se refiere a la celeridad con la que los datos son creados y analizados conforme al crecimiento, y la **Variedad** son las diversas fuentes de donde provienen sin importar el tipo estructurado o no de los datos. (Gil, 2016)

De los anteriores atributos también se derivan tres “uves” más en relación a las propiedades del Big Data.

Conforme Gil (2016) lo menciona, la **Veracidad** es el nivel de fiabilidad o calidad de los datos, la **Visualización** funciona como la representación del comportamiento o traducción grafica de los datos para comprenderlos y tomar decisiones en conformidad, y al **Valor** lo refiere como las acciones realizadas en base al

conocimiento obtenido por la información generada por el procesamiento correcto de los grandes volúmenes de datos.

AREAS DE BIG DATA

La **recolección de datos** es una de las disciplinas de big data, esto es debido a que los datos son generados en grandes volúmenes, y son provenientes de muchas fuentes y de diversos dispositivos distribuidos por todo el mundo que trasmiten, procesan y recolectan los datos que son generados por las diversas actividades como la información generada por las redes sociales, plataformas digitales, datos de geolocalización, entre muchos otros. (Sánchez, D.F, 2018, p.13)

La información se ha convertido en una materia prima de gran valor. El **almacenamiento** masivo de datos y las nuevas fuentes de obtención de los mismos

Un **data warehouse** es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso. es una arquitectura de almacenamiento de datos que le brinda a las empresas la capacidad de comprender y utilizar sus datos para tomar decisiones estratégicas. (Sánchez, D.F, 2018, p.14).

1.3 **Análisis y gestión de procesos orientados a la experiencia del cliente**

Cada vez más se observan los esfuerzos orientados a adecuar las organizaciones al complejo escenario en que se mueven. Cambios de reglas de juego, incremento de la competencia, apertura al mundo a través de la tecnología, hacen al cliente mucho más exigente, modificando sus demandas y necesidades. La Gestión basada en los Procesos, surge como un enfoque que centra la atención sobre las actividades de la organización, para optimizarlas. En este trabajo se considerará a la organización como una red de procesos interrelacionados o interconectados, donde la estructura organizativa vertical clásica, eficiente a nivel de funciones, se orienta hacia una concepción horizontal, desplazándose el centro de interés desde las estructuras hacia los procesos, como metodología para mejorar el rendimiento, concentrándose en el diseño disciplinado y cuidadosa ejecución de todos los procesos de una organización. Concluyendo, la metodología de aplicación de la Gestión basada en Procesos se trata de una herramienta de gestión adecuada para el momento actual, constituyéndose con fuerza como una alternativa exitosa para la obtención de resultados cada vez mejores.

La experiencia del cliente (CX por sus siglas en inglés) se ha convertido en el foco de atención de todas las marcas en los últimos años, ya que las organizaciones han notado la importancia que tiene esto para mantener a los consumidores satisfechos y leales a la marca.

La experiencia del cliente es el conjunto de experiencias que tiene un cliente con una marca durante el tiempo en el que haya una relación de servicio. Mientras mejor

sea la experiencia del cliente, mayor es la probabilidad de incrementar la retención de los clientes, generar nuevos leads, incrementar el ticket promedio, etc.

Bajo este enfoque, la estructura organizativa vertical clásica, eficiente a nivel de Funciones, se orienta hacia estructuras de tipo horizontal, tal cual lo define Ostroff (2000) quien sostiene que no hay contraposición entre modelos, y que cada empresa debe buscar su equilibrio en función de sus propias necesidades y posibilidades.

Revisión de la literatura

Empresas y organizaciones están experimentando un proceso que Vandermerwe y Rada (1988) denominan "servitización", una estrategia basada en la diferenciación en servicios. Romero Amado (2014) apunta a que esta situación obedece a un proceso de desindustrialización, debido al crecimiento económico, la innovación tecnológica y la demanda de servicios en los que el acceso al beneficio es más importante que la propiedad del producto.

Esto supone una estrategia que emplean las empresas para competir en el mercado, y asegurar su sostenibilidad y éxito financiero a largo plazo, la cual está basada en la diferenciación en servicios como oferta central y donde los clientes ven a la organización como proveedora de servicios (Gebauer, Gustafsson, & Witell, 2011, p. 1272). Es decir, en una estrategia que se desarrolla como ventaja competitiva para mejorar el desempeño organizacional (Bozkurt & Kemer, 2014), que, de acuerdo a Gannon, Lynch y Harrington (2014), implica una ejecución impecable para crear o hallar la ventaja -si existe o es posible conseguirla-.

La diferenciación de servicios puede ser contemplada como fuente de ventajas competitivas cuando se considera un proceso donde se combinan personas, ideas habilidades y características únicas entendidas como recursos especiales y de difícil imitación (Bharadwaj, Varadarajan, & Fahy,1993, p. 83; Oliveira Teixeira & Werther Jr., 2013). Así entonces, la fuente de ventaja competitiva es multifactorial y puede venir de cualquier combinación de elementos difícilmente imitables, e ideas que requieren inversión y sinergia por parte de la organización para ser desarrolladas (Dierickx & Cool,1989; Zaridis, 2009).

Modelos de estructuras organizacionales para la gestión del servicio

La estructura organizacional de cada empresa es única por la cultura, el mercado en el que compite y la visión de su directiva (Csaszar, 2012). Sin embargo, a partir de prácticas de prestación de servicios se pueden rastrear elementos estructurales y encontrar puntos de convergencia que permitan caracterizar modelos de servicios.

División entre creación de productos y prestación de servicios

Según Gebauer, Pütz, Fischer y Fleisch (2009, p. 109), el crecimiento en servicios parece requerir una organización separada y distinguible, porque el negocio del servicio debe controlar completamente la identificación y desarrollo de los clientes,

el precio y la entrega de las ofertas. Es decir, las organizaciones están concentradas en una función productiva o manufacturera y añaden ofertas de servicios complementarias, no la inclusión del área de oferta central, lo cual sucede por no entender particularidades de la entrega de servicios y, más bien, centrarse en atender quejas, reclamos o garantías que no generan ingresos.

Gestión por procesos

Para realizar mediciones internas una alternativa completa es la implementación del CRM (Customer Relationship Management) que más que un software es la creación de una cultura corporativa que trabaja poniendo al cliente como prioridad, para conocerlo, entenderlo, atenderlo y satisfacerlo, a través de un equilibrio entre, los procesos, la tecnología, el recurso humano, y el uso eficaz y eficiente de los demás recursos asignados (Universidad Militar Nueva Granada, 2010, pág. 53).

El resultado final de esta estrategia CRM debe ser adquirir nuevos clientes, fidelizar a los clientes a través de la atención oportuna de sus necesidades, afianzar la relación con los clientes actuales al exceder sus expectativas, retener a aquellos clientes que por errores en la cadena de proceso o por factores externos desean retirarse de la compañía.

El CRM debería permitir a la organización conocer aspectos de su cliente que le ofrezcan información demográfica, económica y emocional, segmentar su base de clientes para mayor precisión al momento de ofertar nuevos productos, de fidelizar y de retener, conocer las insatisfacciones de sus clientes para atacar la causa raíz de sus problemas con la organización y crear un modelo único de interacción que convierta a sus clientes en algo más que clientes.

Enfoque Basado En Procesos

Cuando se habla de eficacia hacemos referencia al logro de los resultados que se esperaban tras el establecimiento de los objetivos de la empresa, por otra parte, la eficiencia se orienta al uso adecuado y la optimización de los recursos para el cumplimiento de dichos resultados (Universidad Militar Nueva Granada, 2013, pág. 35); si dentro de la organización no se logra que los procesos se realicen bajo estos dos pilares, podemos decir que se está conduciendo de manera inadecuada el rumbo de la organización.

¿Por qué son importantes los procesos? Si evaluamos el funcionamiento de una empresa, son múltiples las actividades que deben realizarse para la entrega del producto o servicio al consumidor final y más aún para lograr en él una experiencia de satisfacción total, si bien las empresas en el pasado se preocupaban solo por vender un producto terminando con estándares básicos de calidad, los aspectos de competencia, costos, clientes entre otros de la actualidad no permiten a las empresas de hoy conformarse con un producto de baja calidad, ni mucho menos a los clientes aceptarlo. (James F. Riley, 2001, pág. 6.1)

Mejora Continua

La mejora continua, bajo el concepto definido en la norma ISO 9000 tiene por objetivo aumentar la satisfacción de los clientes y de otras partes interesadas, la mejora debe ser una actividad continua al interior de la organización y deriva de la información obtenida de clientes, auditorias, revisión del SGC, que suministran el insumo de todo aquello que tiene una oportunidad de ser mejorado para el bien del cliente y de la empresa. (Organización Internacional de Normalización, 2005, pág. 6)

Elementos del proceso

Los elementos que conforman un proceso son:

- Inputs: recursos a transformar, materiales a procesar, personas a formar, informaciones a procesar, conocimientos a elaborar y sistematizar, etc.
- Recursos o factores que transforman: actúan sobre los inputs a transformar.
    - Factores dispositivos humanos: planifican, organizan, dirigen y controlan las operaciones. La Gestión por Procesos: Un Enfoque de Gestión Eficiente
    - Factores de apoyo: infraestructura tecnológica como hardware, programas de software, computadoras, etc.
- Flujo real de procesamiento o transformación: La transformación puede ser física, de lugar, pero también puede modificarse una estructura jurídica de propiedad.

1.4 **Estrategia CRM como base de Negocio**

Según Berry (1983, p. 25), el marketing relacional “consiste en atraer, mantener e intensificar las relaciones con los clientes” esto puede ser considerado como antecedente y origen del concepto de CRM. Se observa que el CRM constituye una estrategia o modelo de negocio centrado en el cliente, que debe integrar a toda la organización, alineando las distintas funciones que existen con un objetivo común. Su objetivo principal es generar valor para el cliente mediante el conocimiento de sus necesidades o preferencias y mediante la adaptación y personalización de su oferta.

Con respecto a los sistemas de evaluación del CRM se resalta la importancia de establecer un sistema de medida multidimensional que recoja tanto la perspectiva del cliente como las mejoras financieras que puedan derivarse de la estrategia se encuentran los siguientes: un mayor conocimiento del cliente, un aumento de la satisfacción y lealtad de los clientes, una segmentación de mercados y un aumento de las ventas. En este sentido, las variables más destacadas han sido el componente tecnológico, los aspectos organizativos entre los que destaca el apoyo de la alta dirección, la orientación al mercado y al cliente y las prácticas de gestión del conocimiento.

Como factores Tecnológicos según Hansotia (2002, p. 67) citado por Garrillo y Padilla (2010) los sistemas de software CRM permiten a las empresas ofrecer un servicio personalizado, de mayor calidad y a un coste inferior, por lo que la mayor parte de las actividades que generan una orientación cliente-céntrica no serían posibles sin la tecnología adecuada para medir los resultados del CRM, se ha señalado la dificultad de utilizar un único indicador, por lo que en la mayor parte de los modelos se utilizaba una doble escala de medida, que contemplaba tanto resultados financieros como de mercado.

No es cuestión de tratar sólo con clientes que quieren gastar dinero ya sea de forma grupal o individual, tampoco se trata de clientes que son víctimas de nimiedades y meras faltas de educación, por el contrario, se trata de clientes que de una u otra manera tienen que soportar tanto incompetencias, así como mala gestión de todo tipo, que en definitiva tendrán como resultado final una ausencia total de un servicio satisfactorio. (Berry, L., 2007).

El CRM engloba tanto la estrategia como los procesos que comprenden la adquisición, retención y asociación con determinados clientes con objeto de crear un valor superior tanto para la compañía como para el propio cliente, como conjunto de estrategias de negocio, marketing, comunicación e infraestructuras tecnológicas diseñadas con el objeto de construir una relación duradera con los clientes, identificando, comprendiendo y satisfaciendo sus necesidades

El CRM no es sólo una aplicación tecnológica, es una estrategia de negocio que aglutina las funciones de marketing, ventas, servicio al cliente, operaciones, recursos humanos, I+D, finanzas y TI con el objeto de maximizar la rentabilidad de las interacciones con clientes

Supone una integración en toda la empresa de tecnologías trabajando conjuntamente como son almacenamiento de datos, sitio web, intranet-extranet, sistema de apoyo telefónico, contabilidad, marketing, ventas y producción, para permitir la comunicación entre las distintas partes de la organización y así servir mejor a la clientela

Constituye una estrategia de negocio que permite la integración consistente de todas las áreas de negocio que se relacionan con clientes: marketing, ventas, servicio al cliente, mediante una gestión integrada de personas, procesos y tecnología

Conjunto de estrategias que tienen la intención de buscar, recopilar y almacenar la información adecuada, validarla y compartirla a través de toda la organización, con objeto de que después sea utilizada por todos los niveles organizativos para crear experiencias únicas y personalizadas a sus clientes

El CRM no es sólo un paquete de software, sino un enfoque estratégico integral para gestionar la evolución de las relaciones con los clientes que requiere una adaptación continua en respuesta a las necesidades cambiantes del mercado.

1.5 Big Data Management

Big data

Para empezar, se muestran varias definiciones sobre el concepto Big Data, del cual hay incontables definiciones, entre ellas se poseen:

Según (thinkupapp, 2012.), el concepto aplica a la información que no podría ser procesada o analizada por medio de procesos clásicos. Para, Big Data son "porciones masivas de datos que se acumulan con la era que son difíciles de examinar y manejar usando herramientas comunes de administración de bases de datos", y para, Big Data hace referencia "al procedimiento y estudio de gigantes repositorios de datos, tan desproporcionadamente enormes que resulta imposible tratarlos con los instrumentos de bases de datos y analíticas convencionales".

Según (Zdnet.com, Big Data (2010)), Big Data "se refiere a las herramientas, los procesos y procedimientos que permitan a una organización crear, manipular y gestionar conjuntos de datos muy grandes y las instalaciones de almacenamiento".

Dimensiones de Big Data

Existen tres características o dimensiones: Volumen, Velocidad y Variedad.

1) Volumen: "cada día, las empresas registran un aumento significativo de sus datos (terabytes, petabytes y exabytes), creados por personas y máquinas. En el año 2000 se generaron 800.000 petabytes (PB), de datos almacenados y se espera que esta cifra alcance los 35 zettabytes (ZB) en el 2020. Las redes sociales también generan datos, es el caso de Twitter, que por sí sola genera más de 7 terabytes (TB) diariamente, y de Facebook, 10 TB de datos cada día. Algunas empresas generan terabytes de datos cada hora de cada día del año, es decir, las empresas están inundadas de datos." (IBM Big Data and analytics platform, 2012)

***2) Variedad*:** Se puede nombrar que va bastante de la mano con el volumen, puesto que según con éste y con el desarrollo de la tecnología, hay muchas maneras de

representar los datos; es la situación de datos estructurados y no estructurados; dichos últimos son los que se crean a partir de páginas web, archivos de búsquedas, redes sociales, foros, correos electrónicos o producto de sensores en diferentes ocupaciones de los individuos; un caso muestra, “es el transformar 350 mil millones de lecturas de los medidores por año para presagiar el consumo de energía.” (IBM.com, 2012)

3) Velocidad: Tiene relación con la rapidez con que se generan los datos, que es el tamaño en que incrementan los productos de desarrollos de programa (páginas web, archivos de búsquedas, redes sociales, foros, correos electrónicos, entre otros).

Las tres características tienen coherencia entre sí; por ejemplo, analizar 500 millones de registros de llamadas al día en tiempo real para predecir la pérdida de clientes. (IBM Big Data and analytics platform*,* 2012)

El Big Data crece diariamente, como ya se dijo, y una de las justificaciones es que los datos provienen de extensa variedad de fuentes, como por ejemplo la Web, bases de datos, rastros de clics, redes sociales, Call Center, datos geoespaciales, datos semiestructurados (XML, RSS), provenientes de audio y clip de video, los datos causados por los termómetros, datos de navegación de sitios web a lo largo de cierto tiempo, las RFID (Radio Frequency Identification -identificación por radiofrecuencia).(es.wikipedia.org, RFID*, 2010*)

Para el manejo de datos es necesario tener dos componentes básicos, tanto el hardware como el software; respecto al primero, se tienen tecnologías tales como arquitecturas de Procesamiento Paralelo Masivo (MPP), que ayudan de forma rápida a su procesamiento. Para el manejo de datos no estructurados o semiestructurados es necesario acudir a otras tecnologías; es aquí donde aparecen nuevas técnicas y tecnologías, como MapReduce o Hadoop, diseñado para el manejo de información estructurada, no estructurada o semiestructurada.

Las organizaciones que saben sacar beneficio del Big Data tienen la posibilidad de mejorar su táctica y de esta forma quedar en el mercado posicionadas, puesto que

va a hacer uso de nuevos conocimientos, con el enorme volumen de datos o información que maneja a diario, que al principio no se les entregó la suficiente trascendencia, por no tener un instrumento tecnológica que permitiera procesarla. Con la tecnología de Big Data, las organizaciones tienen la posibilidad de dar superiores productos, desarrollar excelentes interacciones con sus consumidores, además, se convierten en más ágiles y competitivas. (Big Data transforms Business, 2012).

Es importante tener en cuenta algunos pasos para la implementación de Big Data, (F. Carrasco, 2013):

- Entender el negocio y los datos. Este primer paso pide un análisis detallado con las personas que hoy laboran y entienden los procesos y los datos que la empresa maneja.
- El segundo paso consiste en determinar los problemas y cómo los datos pueden ayudar. Al momento de conocer los procesos es muy posible que se encuentren los problemas de la empresa o del negocio.
- Establecer expectativas razonables, es decir, definir metas alcanzables; esto se puede lograr si al implementar la solución de un problema éste no presenta alguna mejora, y se debe buscar otra solución.
- Existe una recomendación especial, y es que cuando se inicia un proyecto de Big Data es necesario trabajar en paralelo con el sistema que hoy está funcionando.
- Al intentar de llevar a cabo un plan de Big Data se debería ser flexible con la metodología y los instrumentos; esto se debería a que ambas anteriores son actuales y tienen la posibilidad de llegar a exponer inconvenientes al implementarlas. Esto se puede resolver llevando a cabo indagación e inversión en esta clase de tecnología.
- Es fundamental conservar el propósito de Big Data en mente; esto pues el proceso es pesado y pues no es tedioso, máxime una vez que los procedimientos y herramientas que utilizan Big Data para la investigación de

datos todavía tienen la posibilidad de exponer inconvenientes, y la iniciativa es que se mantenga en mente la meta final del plan sin desanimarse rápido.

DATA MANAGEMENT

En la actualidad existe un gran interés organizacional por lograr lo que se ha denominado “gestión del conocimiento”. Esto implica, primeramente, tomar los datos generados en los procesos empresariales y convertirlos en información al agregarles valor mediante procesos de agrupación, clasificación, etc.; para posteriormente convertir esta información en conocimiento, a través de procesos de separación, evaluación, comparación, etc. (Ponjuán, 2006). Por lo tanto, sin la existencia de datos no se llegaría nunca a obtener conocimiento.

El término calidad, en relación con los datos, toma sentido por el hecho de que los datos al igual que los productos y servicios, deben adecuarse al uso que se les pretende dar. El término preciso para el uso en este caso implica que, dentro de cualquier contexto operacional, el dato que va a ser utilizado satisfaga las expectativas de los usuarios de los datos. Dichas expectativas se satisfacen en gran medida si los datos son útiles para lo que estos los necesitan, son fáciles de entender e interpretar, y además son correctos (Loshin, 2001; Redman, 2001).

Esto quiere decir que la calidad de los datos está asociada a un conjunto de dimensiones o atributos que son los que la definen. Un objetivo fundamental de la definición de las dimensiones es poder establecer un lenguaje común y también focalizar los problemas de calidad de los datos y las oportunidades de mejora (Javed y Hussain, 2003; Naveh y Halevy, 2000). Entre las dimensiones más importantes, pues son las más utilizadas y referenciadas están la exactitud, la integridad, la consistencia y la coherencia (Cong et al., 2007; Levy, 2004; Olson, 2002; Redman, 2001; Strong, Lee y Wang, 1997), es conveniente señalar que éstas deben ser definidas teniendo en cuenta las características propias de cada sector (Gendron y D’Onofrio, 2001).

En la actualidad la tecnología informática es ampliamente utilizada con el objetivo de mejorar el desempeño organizacional en cuanto a calidad de datos, por lo cual

se han desarrollado una amplia gama de softwares. En general, con estos softwares lo que se trata es de realizar un proceso denominado limpieza de datos (data cleaning) (López y Pérez, 2002; Loshin, 2001; Rahm y Hong, 2000). La limpieza de datos implica la exploración en el conjunto de datos, seguida de la validación y verificación del contenido mediante parámetros de validez lógicos lo que permite detectar los posibles problemas y trabajar en su corrección (López y Pérez, 2002). Dentro del proceso de limpieza de datos se desarrollan diversos pasos importantes, los cuales agregan valor en sí mismo al esfuerzo desarrollado por mejorar la calidad. Estos pasos son: el análisis, corrección y estandarización de los datos; la comparación entre datos de diversas fuentes, y la posterior consolidación de los datos en una única base.

En aquellas organizaciones que presentan problemas de calidad de datos, no se emprenden iniciativas para la mejora de esta, motivado en muchas ocasiones por la dificultad de medir la ganancia esperada de éstas. No obstante, la ganancia en este caso siempre se puede materializar en el ahorro de costos (costos relacionados con los diversos aspectos comentados en este artículo) que pueda tener la empresa al poner en práctica la iniciativa de calidad de datos. En última instancia, si al poner en práctica la iniciativa, se espera mejorar el servicio al cliente y en definitiva las relaciones con éstos, pues estas serán razones suficientes para aplicar la mejora proyectada.

1.6 Integrar y Conocer las fases de *big data* y aspectos Legales.

Integrar y conocer las fases de big data

El concepto de Big Data se aplica a toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad específica, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos. (Paredes, 2015).

Segun Zikopoulos (2012), se pueden caracterizar las dimensiones de Big Data con las llamadas cuatro V:

Volumen: La cantidad de datos. Al ser quizá la característica que se asocia con mayor frecuencia a Big Data, el volumen hace referencia a las cantidades masivas de datos que las organizaciones intentan aprovechar para mejorar la toma de

decisiones en toda la empresa. Los volúmenes de datos continúan aumentado a un ritmo sin precedentes.

Variedad: Diferentes tipos y fuentes de datos. La variedad tiene que ver con gestionar la complejidad de múltiples tipos de datos, incluidos los datos estructurados, semiestructurados y no estructurados. Las organizaciones necesitan integrar y analizar datos de un complejo abanico de fuentes de información tanto tradicional como no tradicional procedentes tanto de dentro como de fuera de la empresa. Con la profusión de sensores, dispositivos inteligentes y tecnologías de colaboración social, los datos que se generan presentan innumerables formas entre las que se incluyen texto, datos web, tuits, datos de sensores, audio, vídeo, secuencias de clic, archivos de registro y mucho más.

Velocidad: Los datos en movimiento. La velocidad a la q-ue se crean, procesa y analizan los datos continúa aumentando. Contribuir a una mayor velocidad es la naturaleza en tiempo real de la creación de datos, así como la necesidad de incorporar datos en streaming a los procesos de negocio y la toma de decisiones. La velocidad afecta a la latencia: el tiempo de espera entre el momento en el que se crean los datos, el momento en el que se captan y el momento en el que están accesibles. Hoy en día, los datos se generan de forma continua a una velocidad a la que a los sistemas tradicionales les resulta imposible captarlos, almacenarlos y analizarlos. Para los procesos en los que el tiempo resulta fundamental, tales como la detección de fraude en tiempo real o el marketing “instantáneo” multicanal, ciertos tipos de datos deben analizarse en tiempo real para que resulten útiles para el negocio.

Veracidad: La incertidumbre de los datos. La veracidad hace referencia al nivel de fiabilidad asociado a ciertos tipos de datos. Esforzarse por conseguir unos datos de alta calidad es un requisito importante y un reto fundamental de Big Data, pero incluso los mejores métodos de limpieza de datos no pueden eliminar la imprevisibilidad inherente de algunos datos, como el tiempo, la economía o las futuras decisiones de compra de un cliente.

Big Data comprende un gran volumen y variedad de datos que requieren ser procesados a una alta velocidad para mejorar la toma de decisiones de las empresas, su optimización y mejora continua. El ciclo de vida se puede dividir en las siguientes etapas:

- **Fase de adquisición**. Corresponde a los datos provenientes de fuentes internas y externas (de propiedad de la empresa o adquiridos) en una variedad de formatos y fuentes que permiten satisfacer las necesidades empresariales identificadas. Las fuentes pueden ser de dos tipos: huellas digitales generadas por los humanos y los datos de máquina. Las huellas digitales generadas por los humanos pueden ser un clic, una interacción, un

post en algún foro o red social, un *e-mail*, entre otros. Al contrario, los datos de máquina provienen principalmente de sensores, radares, satélites, cámaras de vídeo, *routers*, servidores, entre otros. En esta fase puede ser necesario el empleo de un ESB (*enterprise service bus*) o un *framework* de integración.

- **Fase de almacenamiento.** Esta fase permite, de forma general, la agregación, consolidación, control de calidad, persistencia y mantenimiento de los datos de la empresa. Estos procesos son conocidos como el «gobierno de los datos». Un estándar bien conocido para implementar este gobierno se denomina MDM (*master data management*).
- **Fase de procesamiento.** Para obtener los resultados esperados, esta fase se divide en subfases, las mismas que conforman el ciclo de vida del procesamiento. Estas subfases son las siguientes: identificar el problema, seleccionar datos y preprocesar los datos (Prajapati, 2013).
  - La subfase «identificar el problema» permite a la empresa identificar ciertos aspectos importantes para su mejor rendimiento.
  - En la subfase «seleccionar datos», se deben seleccionar las fuentes de los datos relacionados con el problema en cuestión y también se deben especificar los atributos de los datos necesarios. Para continuar con el ejemplo anterior, para crear el algoritmo de predicción del clima, las fuentes de los datos serán las estaciones meteorológicas automáticas y el único atributo requerido será el dato correcto.
  - Una vez ya seleccionados los datos, la subfase «preprocesar datos» aplica operaciones sobre los datos, como limpieza, agregación, clasificación y formato.
- **Fase de análisis.** En esta fase se aplican técnicas de minería de datos y los algoritmos de aprendizaje de máquina sobre los datos. Los algoritmos de aprendizaje de máquina más utilizados son los siguientes: regresión, clasificación, agrupación y recomendación. (Almeida M *et.al;* 2015).

Aspectos legales de big data

Las empresas incluyen como procesos recopilar y procesar datos, con el fin de analizar los comportamientos de sus clientes; de esta manera, se pueden personalizar los productos, las ofertas y las promociones, ofreciendo aparentemente beneficios a una población específica de clientes fidelizados o potenciales. En este escenario, *big data* presenta oportunidades para las empresas, por cuanto les permite ser más creativas y competitivas a partir del análisis descriptivo y predictivo de datos, que ayudan a conocer el comportamiento de los clientes a partir de datos históricos que estos mismos han ido acumulando a partir de sus compras de bienes o servicios.

La protección de los datos personales es primordial durante el proceso de recopilación, procesamiento y análisis de información, debido a que la información

básica de los clientes debe ser anónima, para prevenir discriminación y abuso durante el mercadeo digital (por ejemplo, enviar datos básicos tales como nombres, direcciones y números de documento a usuarios de manera errónea, poniendo en riesgo la privacidad). En el ámbito mundial se han establecido regulaciones, leyes y acuerdos que ayudan a garantizar el manejo de datos personales y la privacidad. (Becerra, J. et.al; 2018).

El fenó-meno del internet de las cosas, el *cloud computing* o el *big data* son tecnologías disruptivas, que están revolucionando la forma en la que funciona nuestro mundo.

Y con estas tecnologías, los datos se convierten en el activo más preciado. Hoy en día se crean más datos que nunca en la histo-ria, y recogerlos, almacenarlos y tratarlos es posible de forma más sencilla que nunca. Desde las redes sociales, las compras con tarjetas, las llamadas telefónicas y tantos otros gestos cotidianos ge-neran datos, cuyo estudio es una fuente de valor incalculable.

La expansión de avances como el *big data* conlleva riesgos para la privacidad de los individuos. Así, si los sistemas de análisis de datos son utilizados para un fin que trasciende la legalidad, la privacidad puede resultar quebrada. Es por ello que las organizaciones que quieren to-mar la delantera de los avances analíticos también deberían dar un paso atrás y reconsiderar el diseño que realizan de sus invenciones tec-nológicas, para tomar en cuenta la seguridad de los datos y la privaci-dad ya desde el diseño de la arquitectura de la tecnología, el diseño de sistemas y los procedimientos operativos. Los principios de privacidad deben ser insertados en el código de funcionamiento del dispositivo.

Los riesgos que los avances suponen para la privacidad han hecho cambiar el foco. Ya no debe servir únicamente con cumplir con los principios de protección de datos, ahora también debemos tener en cuenta la privacidad desde el mismo momento en que la tecnología se está diseñando. Es decir, la privacidad debe dejar de ser un mero con-cepto legal para ser una prioridad de negocio. Tomando en cuenta la privacidad ya desde el primer momento, los diseñadores pueden desa-rrollar herramientas que aseguren un mayor grado de protección.

Es posible que privacidad y protección de datos no sean compati-bles con el *big data*, y que haya llegado el momento de reconfigurar los riesgos que esto puede suponer para las personas y los beneficios que el tratamiento de datos puede traernos.

Ciertamente, en el marco legal europeo, la protección de datos y la privacidad se derivan de un derecho fundamental cuya protec-ción no puede obviarse, tal y como ha quedado expuesto en este mismo trabajo. (Gil E. 2015).

Preguntas referentes a la unidad

1.- Se define _____ como: una unidad en sí que cumple un objetivo completo, un ciclo de actividad que se inicia y termina con un cliente o un usuario interno.

2.- Según Zikopoulos, pertenecen a las 4 v características del big data

3.- La limpieza de datos implica la exploración en el conjunto de datos, seguida de la validación y verificación del contenido: verdad o falso

4.- Big Data es el concepto que aplica a la información que no podría ser procesada o analizada por medio de procesos clásicos. Verdadero o falso

5.- Fase del ciclo de vida de big data donde se aplican técnicas de minería de datos y los algoritmos de aprendizaje de máquina sobre los datos.

6.- Big Data comprende un gran volumen y variedad de datos que requieren ser procesados a una alta velocidad para mejorar la _____ de las empresas, su optimización y mejora continua.

7.- La protección de los datos personales es primordial durante el proceso de recopilación, procesamiento y análisis de información, debido a que la información básica de los clientes debe ser _____.

8.- _____ se aplica a toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales.

9.- Atendiendo a los datos de crecimiento exponencial de las herramientas CRM en los últimos años, la causa fundamental es:

10.- El entorno próximo a una empresa este compuesto por factores como:

Respuestas: 1.- Proceso, 2.- Volumen, Veracidad, Variedad, Velocidad, 3.- Verdadero, 4.- Verdadero, 5.- Análisis, 6.- Toma de decisiones, 7.- Anónima, 8.- Big Data, 9.- El avance de las nuevas tecnologías, que proporcionan herramientas eficaces de catalogación, acceso y manejo de la información, 10.- Proveedores.

REFERENCIAS

Acuña Bermeo, C. F. (2008). *Diseño, propuesta de mejora y estandarización de los procesos orientados al cliente de la empresa Edinun* (Bachelor's thesis, QUITO/EPN/2008).

Bernal Zipa, M. M. (2015). Gestión por procesos y mejora continua, puntos clave para la satisfacción del cliente.

De Dios, J., Lara, L., Alberto, R., Carvajal, R., y De Dios, D. J. (2014). La administración de la relación de los clientes (crm), una herramienta para crear estrategias competitivas. EPISTEMUS. https://bit.ly/3Irrd8E

Garrido, A., Padilla, A. (2010). *El CRM como estrategia de negocio: desarrollo de un modelo de éxito y análisis empírico en el sector hotelero español.* Dialnet UNIROJA. https://bit.ly/3BLMwj5

Han, J.,& Kamber, M. (2006). Data Mining: Concepts and Techniques. 2nd Edition. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2006. 770p. ISBN: 978-1-55860-901-3

HANSOTIA, B. (2002): "Gearing Up for CRM: Antecedents to Successful Implementation", Journal of Database Management, 10 (2), pp. 121-132.

Martens C. (2006, 23 de octubre). BI at age 7 [on-line]. Computerworld. https://bit.ly/3AXCSsW

Martínez, L., & El Kadi, O. (2019). Logística integral y calidad total, filosofía de gestión organizacional orientadas al cliente. *Revista Arbitrada Interdisciplinaria Koinonía*, *4*(7), 202-232.

Montoya A., Alveiro C., Saavedra B, Ramiro M. (2013) EL CRM COMO HERRAMIENTA PARA EL SERVICIO AL CLIENTE EN LA ORGANIZACIÓN. Redalyc

Niño, M., & Illarramendi, A. (2015). ENTENDIENDO EL BIG DATA : ANTECEDENTES , ORIGEN Y DESARROLLO POSTERIOR. 1–8. Dyna New Technologies. https://doi.org/10.6036/NT7835

Pérez Fernández, J. A. (1999). *Gestión de calidad orientada a los procesos* (No. HF5549 P415)

Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Magazine. Vol.11-5. p.68-70. DOI: http://dx.doi.org/10.1609/aimag.v11i4.873

UNIDAD II. Análisis y Estadística.
OBJETIVO: Describir técnicas de configuración de datos para el análisis mediante algoritmos estadísticos y técnicas de aprendizaje automático, así como la generación de conocimiento en Big Data.

2.1 Técnicas del tratamiento de datos
Análisis e interpretación de la información

Análisis cuantitativo

El análisis consiste en descomponer un todo en cada una de sus partes. Se ha elegido la siguiente definición de análisis de la información: “Es la manipulación de hechos y números para obtener cierta información mediante técnicas que al investigador posteriormente le podrán permitir tomar decisiones (Ortiz, 2010, p. 135). El autor describe como se ejecuta el proceso de análisis de información y su consecuente como es la toma de decisiones en su investigación.

Criterios de ejecutar análisis de información

Se consideran dos criterios como son las características de la muestra y variables de estudio (Ortiz, 2010, p. 247). La primera ubica la realidad de la investigación, es decir, actores, ambientes físicos, contextos, temporalidad que identifica las particularidades del tema. La segunda, describe si lo hallado y evaluado se relacionan entre sí, incide en el cálculo de porcentajes, medidas de tendencia central y variabilidad. Entre las principales formas de análisis de información tenemos:

Estadística descriptiva para cada variable. Se analiza una sola variable separada de las demás, es decir, se identifica, relaciona y contrasta por cada una. Se utiliza cuando se mide por “nivel de intervalo o razón”. (Ramírez 2010, p. 225). Se trabaja con las medidas de tendencia central. Análisis bivariado. Se relacionan categorías de una variable con las categorías de la segunda variable mediante “el uso de tablas de contingencias” (Ramírez 2010, p. 227). Se debe tener en cuenta, según Ramírez (2010, p. 228): “El titulo debe reflejar la información que contiene la tabla. (sic) Incluir un subtítulo para cada columna que se integre a la tabla. Indicar el 100% cuando la tabla se exprese en términos porcentuales. Indicar al final de cada columna el número total de casos o categorías que corresponde”

ANÁLISIS PARAMÉTRICO

Se basa en el análisis de los parámetros estadísticos relacionados a la muestra y la población. Las pruebas paramétricas más utilizadas son:

Tabla 1 : Pruebas paramétricas más utilzadas.

PRUEBA	TIPOS DE HIPÓTESIS
Coeficiente de correlación de Pearson	Correlacional
Regresión Lineal	Correlacional/Causal
Pruebas t	Diferencia de grupos
Contraste de la diferencia de proporciones	Diferencia de grupos
Análisis de varianza	Diferencia de grupos/causal
Análisis de varianza/covarianza	Correlacional/causal

FUENTE: Triola, 2009, p. 386-395

Coeficiente de correlación de Pearson

Es una prueba estadística para analizar la relación entre dos variables medidas en un nivel por intervalos o de razón. “Se simboliza por r”. (Hernández, 2006, p. 230). La correlación de Pearson se define como: “A mayor X, mayor y” “A mayor X menor Y”, “Altos valores en X están asociados con altos valores en Y”. Altos valores en X se asocian con bajos valores en Y” (Hernández, 2006, p. 230)

Figura 1: Análisis sugerido de la r de Pearson

Valor del Coeficiente de Pearson	Grado de Correlación entre las Variables
$r = 0$	Ninguna correlación
$r = 1$	Correlación positiva perfecta
$0 < r < 1$	Correlación positiva
$r = -1$	Correlación negativa perfecta
$-1 < r < 0$	Correlación negativa

FUENTE: Hernández, 2010, p. 312

Ejemplo de correlación r de Pearson (EAP Psicología):

En la siguiente investigación “Insatisfacción de la imagen corporal y autopercepción en adolescentes de una escuela secundaria”

Objetivos. Identificar la insatisfacción de la imagen corporal y la autopercepción en adolescentes de una escuela secundaria de la Ciudad de México. Material y métodos. Diseño. Estudio prospectivo, transversal, analítico. Población. Estudiantes de una secundaria del sureste del Distrito Federal. Muestreo probabilístico en alumnos de 9 a 14 años. Actividades. Se solicitó firma de carta de consentimiento informado y se aplicó cuestionario, Body Shape Questionnaire (BSQ) (Cuestionario sobre Forma Corporal), y de autopercepción

Análisis estadístico

Se utilizó coeficiente de correlación de Pearson cuando las variables fueron medidas a nivel escalar y de Spearman cuando el nivel de medición fue ordinal, así como t de Student de muestras independientes, con nivel de significancia de 0.05. Se utilizó el programa estadístico SPSS versión 17. (Aceves y García, 2011, p. 130)

Ejemplo de análisis de varianza (EAP Psicología)

Para conocer las características psicométricas de la escala, se llevaron a cabo análisis de consistencia interna (alfa de Cronbach) y análisis factorial de componentes principales con rotación Varimax. Se consideraron como válidos valores de correlación ítem–total iguales o mayores a 0.40 y los factores con un mínimo de tres preguntas. Posteriormente, con el objetivo de conocer la validez predictiva de la escala, se realizó un análisis discriminante por el método de inclusión por pasos con la muestra total y dos submuestras tomadas al azar. Para la comparación de medias entre las variables de imagen corporal entre las mujeres que presentan conductas alimentarias de riesgo (CAR) y las que no las presentan se utilizó la prueba t de Student. Posteriormente se utilizó la prueba de análisis de varianza (ANOVA de una vía) y la prueba *post hoc* de Scheffé para observar si existen diferencias entre Imagen Corporal y edad entre las diferentes categorías del IMC. Los datos fueron analizados con el programa SPSS para Windows versión 15.0. (Rodríguez, 2010, p. 328)

Análisis no paramétrico

Evitan el uso de supuestos de la forma de distribución poblacional; aceptan distribuciones *no normales*; las variables no necesariamente están medidas en intervalos o en razones. Pueden analizar datos normales u ordinales.

Tabla 2: PRUEBAS NO PARAMÉTRICAS MÁS UTILIZADAS

PRUEBA NO PARAMETRICA	TIPOS DE HIPÓTESIS
Regresión múltiple	Una dependiente y dos o más independientes
Análisis lineal de patrones	Varias, secuencia causal
Análisis de factores	Varias (intervalos o razón)
MANOVA y correlación canónica	Varias independientes y varias dependientes
Análisis discriminante	Varias independientes (intervalos o razón) y una dependiente (nominal u ordinal)

FUENTE: Ramírez 2010, p. 331

RESUMEN: Numerosos protocolos de investigación en pediatría trabajan con variables de tipo cualitativo, por ejemplo, sexo del recién nacido (masculino, femenino) o grado de desnutrición (leve, moderado, severo).

Para determinar la asociación o independencia de dos variables cualitativas con un cierto grado de significancia, se dispone de una herramienta estadística frecuentemente utilizada, el test de chi–cuadrado ($\chi 2$). El presente artículo explica el fundamento teórico del test, la metodología de cálculo del estadístico $\chi 2$ y su correcta interpretación, ejemplificando estos conceptos mediante una investigación real. En términos simples, el test de chi–cuadrado ($\chi 2$) contrasta los resultados observados en una investigación con un conjunto de resultados teóricos, estos últimos calculados bajo el supuesto que las variables fueran independientes. La diferencia entre los resultados observados y esperados se resume en el valor que adopta el estadístico $\chi 2$, el cual tiene asociado un valor–p, por debajo del cual se acepta o rechaza la hipótesis de independencia de las variables. De esta forma, al someter los resultados de una investigación(sic) test de chi–cuadrado ($\chi 2$) el investigador puede afirmar si dos variables en estudio están asociadas o bien son independientes una de la otra, afirmación que cuenta con un sustento estadístico. (Cerda y Villarroel, 2007, p. 414)

Análisis cualitativo

El análisis surge en base a cada caso o experiencia del investigador. Se han seleccionado dos técnicas relevantes para el análisis cualitativo como son el *análisis del discurso* y la *teoría fundamentada*.

El análisis del discurso

El concepto considera la articulación del texto hablado y escrito. En términos de Van Dijk (2013a, p. 12) “No sólo incluye “análisis”, sino también “teorías”, “aplicaciones”, “crítica” y otras dimensiones de investigación…”. Es decir el uso del lenguaje al ser expresado cuyo fundamento es “el estudio crítico de la reproducción discursiva de la dominación en la sociedad” (Van Dijk, 2013b, p. 151).

RESUMEN: Numerosos protocolos de investigación en pediatría trabajan con variables de tipo cualitativo, por ejemplo, sexo del recién nacido (masculino, femenino) o grado de desnutrición (leve, moderado, severo).

Para determinar la asociación o independencia de dos variables cualitativas con un cierto grado de significancia, se dispone de una herramienta estadística frecuentemente utilizada, el test de chi–cuadrado ($\chi 2$). El presente artículo explica el fundamento teórico del test, la metodología de cálculo del estadístico $\chi 2$ y su correcta interpretación, ejemplificando estos conceptos mediante una investigación real. En términos simples, el test de chi–cuadrado ($\chi 2$) contrasta los resultados observados en una investigación con un conjunto de resultados teóricos, estos últimos calculados bajo el supuesto que las variables fueran independientes. La diferencia entre los resultados observados y esperados se resume en el valor que adopta el estadístico $\chi 2$, el cual tiene asociado un valor–p, por debajo del cual se acepta o rechaza la hipótesis de independencia de las variables. De esta forma, al someter los resultados de una investigacional (sic) test de chi–cuadrado ($\chi 2$) el investigador puede afirmar si dos variables en estudio están asociadas o bien son independientes una de la otra, afirmación que cuenta con un sustento estadístico. (Cerda y Villarroel, 2007, p. 414)

Análisis cualitativo

El análisis surge en base a cada caso o experiencia del investigador. Se han seleccionado dos técnicas relevantes para el análisis cualitativo como son el *análisis del discurso* y la *teoría fundamentada*.

El análisis del discurso

El concepto considera la articulación del texto hablado y escrito. En términos de Van Dijk (2013a, p. 12) “No sólo incluye “análisis”, sino también “teorías”, “aplicaciones”, “crítica” y otras dimensiones de investigación…”. Es decir, el uso del lenguaje al ser expresado cuyo fundamento es “el estudio crítico de la reproducción discursiva de la dominación en la sociedad” (Van Dijk, 2013b, p. 151).

La teoría fundamentada

Es una metodología de análisis que permite identificar las relaciones existentes en el estudio de un discurso o texto. En términos de Strauss (2004, p. 51):" La teoría fundamentada no es una teoría, sino una metodología para descubrir teorías que dormitan en los datos". Las estrategias básicas son la codificación de frases y expresiones, la formación de categorías que hacen referencia al comportamiento social estudiado, la reflexión y análisis sobre estas categorías, así como sus relaciones para intentar obtener un esquema teórico. (Secretaría de Marina Armada de México, 2010, p. 40). Reafirma dicha postura Charmaz (2005, p. 507): "…unas directrices analíticas que permiten a los investigadores focalizar su recolección de datos y construir teorías de rango medio a través de sucesivas recolecciones de datos y desarrollos conceptuales"

Ejemplo de teoría fundamentada: "No me interesaba conocer por qué las adolescentes se embarazan sino conocer las circunstancias en las que se producen estos embarazos, las circunstancias subjetivas, esto es, sus circunstancias. Argumenté que se sabe mucho de la salud reproductiva de los jóvenes, pero muy poco de la salud reproductiva en los jóvenes. Encontré que el embarazo para las jóvenes es un hecho sentimental y biográfico que significa un punto de inflexión en su carrera personal. El contexto de la interacción de la joven que se embaraza es de noviazgo en serio en el que las ideas de amor romántico y las reglas de género guían su comportamiento. Las jóvenes que se embarazan se enamoran de alguien en concreto y no de manera abstracta. Una joven durante la entrevista comentó: "Yo tenía amigos, pues como cualquier persona joven. Sí, me manejaba bien en la casa, hasta que llegó el muchacho, y nos hicimos novios y ahí fue donde empezó todo" (De la Cuesta, 2006, p. 139)

Finamente se puede establecer la siguiente comparación entre ambos análisis de investigación.

Tabla 3: COMPARACIÓN ENTRE EL ANÁLISIS DE LOS RESULTADOS DESDE LA INVESTIGACIÓN CUANTITATIVO Y LA INVESTIGACIÓN CUALITATIVA

	Análisis cuantitativo	Análisis cualitativo
Análisis de los resultados		
Objeto del análisis	La variable (análisis por variable, impersonal)	El individuo (análisis por persona)
Objetivo del análisis	Explicar la variación de las variables.	Comprender a los sujetos.

Uso de técnicas matemáticas y estadísticas	Máximo	Ninguno

Fuente:

LA INTERPRETACIÓN

La interpretación consiste en inferir evidencias a partir de los datos en conjunción con la teoría de referencia y conforme a ciertas reglas: Si los datos constituyen evidencias relevantes, confirmadoras o refutadoras de la hipótesis, o no constituyen evidencias por ser irrelevantes. (Ortiz, 2010). Explicar el contexto de la investigación es facilitado por la interpretación que se efectúa en forma directa a la mayor relación de datos que se relacionen entre sí del estudio ejecutado. (Fernández 2007, p. 65)

Se ha considerado tres criterios:

Primer criterio: Critica, es decir, se evalúa si defiende las ideas del autor. Se afirma las ideas desarrolladas por otro autor. Se considera una afirmación no revisada anteriormente. Se reconsidera una afirmación, superando sus limitaciones, configurándose una explicación aproximada con la teoría manejada. (Ramírez 2010, p. 334)

Segundo criterio: Verificación de hipótesis, es decir, se identifica las consecuencias lógicas y empíricas de las hipótesis con base en el análisis de los referentes de sus variables.

De las hipótesis o estadísticas accesorias, se identifica el soporte teórico suplementario de la hipótesis de trabajo que les corresponda, y de ser posible, someterlas a contrastación empírica independiente. Se recopila datos que sean comparables con los referentes de las variables, y con base en éstos, contrastar las hipótesis empíricas o estadísticas. Se determina el grado en que los datos concuerdan con las hipótesis o que los coeficientes de los datos contrastados contribuyen, esto es la evidencia de la relación entre las variables. (Ortiz, 2010)

Tercer criterio: Contrastación de hipótesis, es decir, la unidad de análisis se refiere al sistema concreto o conjunto que designa el dominio de la relación entre las variables y en el cual se presentan o manifiestan sus consecuencias en términos empíricos o factuales. Si las técnicas de comprobación son los procedimientos empíricos y los instrumentos conceptuales mediante los cuales se lleva a cabo la contrastación, al ser aplicados a una muestra representativa de la unidad de análisis.

Si el marco teórico, se refiere a alguna(s) teoría científica que fundamenta el proceso de comprobación, y que permita proyectar los términos de la contrastación empírica y la interpretación de los resultados que se obtenga de ésta. (Ortiz, 2010)

Facultad de letras y ciencias humanas análisis de los efectos psicosociales de la violencia política en tres distritos del departamento de Huancavelica con distintos niveles de afectación. Tesis para optar el título de Licenciada en Psicología con mención en Psicología Social que presenta la Bachiller: María Eugenia Moyano García

Tipo de soporte social

Los resultados acerca del tipo de soporte social que recibieron las víctimas de los tres distritos estudiados muestran que las manifestaciones de ayuda son diversas. Las más mencionadas son las siguientes: *curarse del daño físico* (5.4%), *camas/frazadas/colchas* (5.4%), *alimentos* (4.7%), *levantar / enterrar el cuerpo del familiar* (4.0%) y el *soporte emocional* (4.0%).

Otras manifestaciones de soporte social encontradas fueron: la *búsqueda/recuperación del familiar* (3.4%), ayuda para *denunciar los hechos* (3.4%) y los *implementos para la cocina y la alimentación* (3.4%). Se ha encontrado, además, que el distrito menos afectado (Julcamarca), recibió más ayuda para la *búsqueda /recuperación del familiar* que el distrito más afectado (Santo Tomás de Pata).

En cuanto a las diferencias de género, se encontró que las mujeres recibieron más *alimentos* y *soporte emocional* que los hombres, aspectos que puede ser explicado a partir de las teorías del rol de género según las cuales, las mujeres y los hombres cumplen distintos roles y son socializados según una cultura particular (Monat et al.,2007). Estos planteamientos sustentarían, por ejemplo, que las mujeres andinas, al encargarse de las labores de la casa (Pinzás, 2001), reciban los alimentos y que, por haber sido socializadas con cierto grado de vulnerabilidad, reciban más soporte emocional que los hombres (Stokes & Wilson, 1984, citado en Monat et al., 2007; Pika, 1998).

Por otro lado, es interesante notar que las necesidades de alimentación y vestido aparecen junto al soporte emocional en el factor *necesidades básicas y soporte emocional;* esto podría explicarse a partir de las características de una cultura colectivista (como la andina), en donde los vínculos son fundamentales para el sí mismo (Triandis, 2001) y en las cuales la salud mental se asocia al soporte social (Velásquez, 2007); en este caso el soporte emocional aparece como un aspecto igual de importante que las necesidades básica para la supervivencia como la alimentación o el vestido.

Respuestas de afrontamiento

Las conductas de afrontamiento más usadas por las víctimas de la violencia fueron el *desplazamiento hacia otros lugares* (47.0%) y *no poner una denuncia* (34.2%). Otras conductas similares, pero con menor porcentaje de mención fueron: *escaparse* (17.4%) y *esconderse de los agresores* (16.1%). Según la literatura revisada, estas conductas se ubicarían dentro de la familia del afrontamiento del escape – huída, cuyas principales funciones en el proceso adaptativo son dejar atrás un entorno inquietante o evitar una acción directa hacia el mismo, pues se presenta como incontrolable; estas formas de afrontamiento se asocian al miedo (Skinner & Zimmer, 2007.).

Los resultados muestran, además, que los pobladores del distrito más afectado (Santo Tomás de Pata) se desplazaron más a otros lugares y se escaparon de los agresores más que los del distrito menos afectado (Julcamarca). Asimismo, se encontró que las mujeres se escaparon de los agresores más que los hombres. Esto puede deberse a los roles de género durante la época de violencia: quizás los hombres asumieron un rol más defensivo, a través de organizaciones como los comités de autodefensa o las rondas campesinas, mientras que las mujeres se quedaron en casa al cuidado de los hijos y la única manera de protegerlos era escapar.

Otras formas de afrontamiento mostradas fueron la *denuncia* (22.8%) y el *dar aviso a autoridades, comunidad, Fuerzas Armadas y Policía* (6.7%). Ambas conductas pueden ser ubicadas dentro de la familia de la delegación, cuya principal función en el proceso adaptativo es la de identificar el límite de recursos con los que se cuenta y coordinar la resistencia, así como los recursos sociales disponibles (Skinner & Zimmer, 2007).

Otra conducta reportada por los participantes fue el *retorno a la comunidad*, la cual podría ubicarse dentro de la familia de adaptación, cuya función en el proceso adaptativo es ajustar flexiblemente las preferencias a las opciones (Skinner & Zimmer, 2007). La mayoría de las familias desplazadas por la violencia no contaban con los medios básicos para la subsistencia (Consejo Económico y Social de las Naciones Unidas, 1996). Esta situación pudo haber influido en que muchas regresaran a sus lugares de origen, donde sí podían alimentarse a través de la actividad agrícola; esto se ajusta a lo señalado respecto a la familia de afrontamiento de la adaptación, según la cual, es preciso elegir entre las opciones disponibles: quedarse en el lugar al que se desplazaron o regresar a su lugar de origen. (Skinner & Zimmer, 2007).

Otras respuestas se encuentran dentro de la familia de la resolución de problemas a través de la creación de una estrategia, una acción instrumental o la planificación, con la finalidad de modificar acciones para que éstas sean más efectivas (Skinner & Zimmer, 2007). Tal es el caso de la *búsqueda del familiar* (13.4%) o de la *recuperación del cuerpo del familiar* (3.4%), en donde hubo la necesidad de planificar una acción como, por ejemplo: acudir a la base contrasubversiva o a la

cárcel, viajar a otro pueblo, rescatar el cuerpo de un barranco, etc. Los resultados mostraron también que, en el distrito menos afectado (Julcamarca), se recuperaron en mayor medida los cuerpos de los familiares fallecidos en relación a Santo Tomás de Pata, el distrito más afectado.

También se podría considerar dentro de esta familia, el *enterrar / velar al familiar* (15.4%), pero solo en los casos en los que hubo que planificar para lograrlo; muchas personas tuvieron que hacer este tipo de ceremonias con mucha rapidez e incluso a escondidas, ya que eran amenazadas por los subversivos. Este tipo de ceremonias también podrían ser ubicadas en la familia de independencia en su forma de regulación / expresión emocional (Skinner & Zimmer, 2007).

La búsqueda de apoyo (6.7%) una de las principales formas de afrontamiento en todas las edades (Skinner & Zimmer, 2007) también apareció, aunque en menor porcentaje que otras formas de afrontamiento; esto quizás se deba a la poca confianza en las autoridades (CVR, 2003) y al miedo, sentimiento generalizado entre las poblaciones que sufren violencia (Ackermann et al., 2005; Crenshaw, 2004; Kimhi & Shamai, 2004; CVR, 2003; Beristain y cols, 1999). Los resultados mostraron, también, que las mujeres solicitaron menos ayuda / apoyo que los hombres; lo cual llama la atención de acuerdo a las teorías basadas en el rol de género que consideran que las mujeres -al estar preparadas para establecer relaciones y ser socializadas con cierto grado de vulnerabilidad a diferencia de los hombres- podrían tener menos dificultades para solicitar apoyo, ya que esto no sería mal visto (Monat et al., 2007); sin embargo esto puede deberse a las características del conflicto en donde los hombres asumieron roles defensivos y necesitaban solicitar apoyo para defenderse de los ataques subversivos.

RESULTADOS.

Se describen, pero principalmente se examinan los resultados encontrados; esto es tratando de articular la información recolectada. Es dar cuenta de la información por variable o categoría de estudio. Además, en esta parte se busca relacionar los datos con la teoría. Esta es la parte más importante de la tesis, en la que “el investigador da cuenta de sus hallazgos, de las relaciones que establece entre los datos y de las relaciones de éstos con la teoría” (Díaz, 2007, Entrevista).

Se deben seleccionar los descriptores y/o los cuantificadores más relevantes y característicos de la muestra. Se han de articular los resultados con las variables de estudio.

Se debe estructurar una síntesis coherente y sustentar las afirmaciones, exponiendo datos concretos. Además, es pertinente consignar los cuadros estadísticos más significativos y presentar gráficos que permitan visualizar los

resultados del análisis de datos: diagrama de barras, pies, histograma o aquellos que se consideren coherentes.

Tabla 4: COMPARACIÓN DE LOS RESULTADOS ENTRE INVESTIGACIÓN CUANTITATIVA Y CUALITATIVA

	Investigación cuantitativa	Investigación cualitativa
RESULTADOS		
Presentación de los datos	Tablas (enfoque relacional)	Fragmentos de entrevistas, textos (enfoque narrativo)
Generalizaciones	Correlaciones. Modelos causales. Leyes. Lógica de la causalidad	Clasificaciones y tipologías. Tipos ideales. Lógica de la clasificación.
Alcance de los resultados	Se busca generalizaciones (inferencias)	Especificidad

Fuente: Corbetta (2007, p. 43)

Ejemplo en presentación de resultados (EAP Ciencias Empresariales)

“Se ha utilizado la prueba *t* de *Student* de diferencia de medias para muestras independientes para analizar la influencia de las variables demográficas y de personalidad en las evaluaciones del vendedor sobre comportamientos de venta éticamente cuestionables (véase la tabla 5). Es decir, la variable a contrastar es la suma aritmética de los seis escenarios planteados descritos en la tabla 2. Las variables agrupación han sido divididas en las siguientes categorías; género (hombres vs. mujeres), edad (vendedores menores 34 años

vs. vendedores mayores de 34 años), formación (vendedores sin estudios universitarios vs. con estudios universitaríos), maquiavelismo (vendedores menos maquiavélicos vs. Más maquiavélicos) y materialismo (vendedores menos materialistas vs. más materialistas). A excepción de la variable formación, el resto fueron dividas por la mediana en consonancia con investigaciones similares (Altemeyer, 2004)” (Román y Rodríguez, 2011, p. 94)

Las correlaciones y la consistencia interna (alfa de Cronbach) de las variables del estudio. Todas las correlaciones encontradas son significativas en el nivel p<.OI. Todos los coeficientes *a* superan el criterio de aceptación de .70 (Bagby, Ryder, SchuUer Grande, Abascal y Marshall, 2004; Kaplan y Sacuzzo 2001, 2006; Nunnally y Berstein, 1994). Los resultados de los análisis diferenciales realizados en función del género muestran que las mujeres presentan un mayor nivel de agotamiento que los hombres, tanto en TI (t= -2.57, p<0,01), como en T2 (t= -3.01; p:<0.01); por otro lado, los hombres presentan un mayor nivel de cinismo que las mujeres, tanto en TI (t= 4.96; p<0.01), como en T2 (t= 3.31; p<0.01). En el caso de las mujeres, los resultados muestran que existen efectos directos positivos significativos tanto del agotamiento en TI *{ß= .71,p<.01),* como del cinismo en T2 *(ß*= .32, *p<.01),* sobre el agotamiento en T2 (que explican el 50% y el 9% de la varianza, respectivamente). Es decir, el nivel de agotamiento en TI predice el agotamiento en T2. Asimismo, el nivel de cinismo predice el cambio producido en el nivel de agotamiento.

Ejemplo en presentación de resultados (EAP Psicología)

La composición de la versión final de la prueba fue de 26 reactivos, integrados en tres factores: El factor I se denominó insatisfacción corporal, consta de 10 reactivos con una media de 3.1 (d.e.=0.8), un total de la varianza explicada de 17.9% y un Alfa de Cronbach de 0.84. El factor II se denominó interiorización del ideal estético de delgadez, se formó por 10 reactivos, con una media de 3.5 (d.e.=0.9), varianza explicada de 15.2% y un Alfa de Cronbach de 0.89. El factor III se denominó influencia social, se formó de seis reactivos, con una media de 3.8 (d.e =0.8), varianza explicada de 9.9% y un alfa de Cronbach de 0.82.

Para observar el nivel de confiabilidad total del instrumento se realizó un análisis de consistencia interna de los 26 reactivos restantes con el cual se obtuvo un valor alfa de Cronbach de 0.94. Se llevó a cabo un análisis de comparación con la prueba t de Student la cual indicó que las mujeres con CAR mostraron mayor insatisfacción corporal, interiorización del ideal estético de la delgadez e influencia social. No se presentaron diferencias estadísticamente significativas respecto a la comparación por edad, mientras que resultaron significativas las comparaciones entre el IMC y los factores de insatisfacción corporal e interiorización del ideal estético de delgadez. (Rodríguez, 2010)

2.2 Técnicas analíticas en Big Data.

El análisis de datos es el estudio exhaustivo de un conjunto de información cuyo objetivo es obtener conclusiones que permitan a una empresa o entidad tomar una decisión.(Westreicher, 2020)

Según Joyanes (2013.) la analítica de Big Data es la utilización de técnicas analíticas avanzadas en conjuntos de Big Data. Por consiguiente, analítica de Big Data se compone de dos teorías: analítica (analytics) y Big Data. Las organizaciones

necesitan recurrir a la analítica de Big Data para tomar decisiones de negocio lo más acertadas posibles. Las herramientas de analítica deben contemplar: reporting, query y visualización, analítica predictiva, analítica Web, analítica social, y social listening, analítica especializada para Big Data procedentes de fuentes M2M o Internet de las cosas, entre otras.

Existe un gran número de herramientas de software propietario y software abierto que soportan Big Data:

REPORTING, QUERY Y VISUALIZACIÓN

Proveedores tales como SAS, IBM (Cognos), SAP (Business Object), Tableau, QlickView y Pentaho tienen buenas soluciones para visualización, query y reporting que ayudan en el análisis de Big Data.

ANALÍTICA PREDICTIVA

SAS e IBM (SPSS) ofrecen herramientas que permiten construir modelos predictivos basados en Big Data. Aquí se destaca R como un paquete de código abierto muy utilizado para análisis estadístico en grandes conjuntos de datos sobre plataformas Hadoop.

ANALÍTICA WEB

Avinash Kaushik, (2012) citado por Joyanes *et al*, (2013) considera que posiblemente la persona más prestigiosa en el mundo de la analítica Web, la define como: “El análisis de datos cualitativos y cuantitativos de su sitio Web y de la competencia, para impulsar una mejora continua de la experiencia online que tienen tanto los clientes habituales como los potenciales y que se traduce en unos resultados esperados (online y offline)”. Las herramientas de analítica Web son imprescindibles en la analítica de Big Data, por el enorme volumen de datos que generan los medios sociales. Estas herramientas, como considera la definición de Kaushik, deberán integrar los datos sociales con la información de la competencia y la información fuera de línea, de modo que proporcionen una visión completa del comportamiento de cada visitante en el tiempo, y también en los diferentes canales.

El propio Kaushik recomienda algunas herramientas para el trabajo diario: Omniture (hoy de Adobe, Adobe Digital Marketing), Coremetrics (hoy de IBM), y Webtrends, entre las herramientas propietarias, y entre las herramientas gratuitas: Google Analytics y Yahoo Analytics.

ANALÍTICA SOCIAL.

Analítica social para Lovett, (2011) citado por Joyanes *et al*, (2013) es la disciplina que ayuda a las empresas a analizar, calcular y explicar el rendimiento de las iniciativas de social media en el contexto de objetivos empresariales específicos. Esto significa que ayuda a entender cómo las personas perciben su marca y cómo responden a productos corporativos, servicios y mensajes, preferentemente de

marketing. La analítica social mide los resultados de una estrategia en medios sociales. En todo proyecto de social media debe estar siempre presente la escucha social: escuchar antes de poner en marcha la estrategia social, escuchar durante el desarrollo de la estrategia social.

La escucha social busca saber lo que se está hablando de su marca. Se trata de analizar las conversaciones que se dan sobre las marcas entre sus usuarios. Las fuentes que maneja un plan de escucha social son muy variadas y van desde Facebook y Twitter hasta foros de discusión, blogs… En esencia, el objetivo es analizar las conversaciones que se dan en un primer nivel entre los consumidores. Se trata de escuchar a los clientes y medir su compromiso (engagement) con la marca. Proveedores de herramientas de escucha social hay muchos, pero destacaremos Radian6 de Salesforce.com, Lithium y Attensity, como empresas del mundo de social media, pero también los grandes proveedores de software han sacado al mercado herramientas específicas como es el caso de Oracle con Collective Intellect, SAS con Social Media Analytics y IBM con Cognos Consumer Insight.

ANALÍTICA M2M

La analítica de datos entre máquinas (M2M) y de Internet de las cosas requiere una analítica especializada debido precisamente a las características particulares de las fuentes de datos de donde proceden. La explosión de los datos máquina a máquina (M2M) mediante sensores inalámbricos se está volviendo un elemento común en diferentes dispositivos industriales y para el consumidor, como máquinas expendedoras, productos para atención médica, sistemas de seguridad para hogares, parquímetros y automóviles. También están cada vez más omnipresentes en la industria del transporte: por ejemplo, los trenes de alta velocidad de Japón tienen sensores que verifican la actividad sísmica, cambios ambientales, tráfico inesperado en las vías y otras anomalías. Una herramienta muy conocida y de gran utilidad en el manejo de Big Data es Splunk, un software para buscar, monitorizar y analizar datos generados por máquinas por aplicaciones, sistemas e infraestructuras de TI vía interfaces o en registros log de las redes. (Lovett, (2011) citado por Joyanes *et al*, (2013)

PLATAFORMAS DE ANALÍTICA DE BIG DATA

Un informe del TDWI (The Data Warehousing Institute TM) selecciona los proveedores de soluciones de analítica de Big Data que considera de mayor relevancia técnica, en el campo profesional, aunque la oferta es mucho más amplia:

- Cloudera
- EMC Greenplum
- IBM
- Impetus Technologies

- Kognitio
- ParAccel
- SAP
- SAND Technology
- SAP
- SAS
- Tableau Software
- Teradata

CLOUD COMPUTING

Numerosas organizaciones van mirando a la nube (cloud computing), y desde hace un par de años están diseñando estrategias para su migración. Las dificultades son grandes por los diferentes tipos de nubes y modelos de despliegue existentes, pero la decisión poco a poco va siendo tomada por las direcciones de las empresas, esencialmente, por la facilidad de despliegue, flexibilidad, ahorro de costes… A medida que se produce esta migración también se plantea la necesidad de explotar las nuevas tendencias de Big Data. Es necesario, en la toma de decisiones, definir una estrategia de desarrollo e integración de Big Data en los entornos de la nube. Cada día más, numerosos proveedores ofrecen plataformas de Big Data, en la nube. Amazon Web Services de Amazon, proveedor líder en cloud computing, ofrece una marco de trabajo Hadoop integrado en su servicio Amazon Elastic MapReduce; Google con su servicio Google Cloud Platorm permite a las organizaciones construir aplicaciones, almacenar grandes volúmenes de datos, y analizar estos grandes datos; EMC, el gran fabricante de soluciones de almacenamiento, ofrece sus herramientas en torno a GreenPlum que facilita la integración de Big Data y cloud; Fujitsu, el fabricante de hardware de grandes máquinas; TrendMicro, la compañía de seguridad, ofrecen soluciones para la integración entre la nube y los grandes volúmenes de datos. El tema de la integración de Big Data y cloud computing es un tema candente en 2013, y lo seguirá siendo en los próximos años. Una prueba de su actualidad, lo da el NIST de los Estados Unidos, referencia obligada en normas y estándares de Tecnologías de la Información, que inició el año con un workshop “NIST Joint Cloud and Big Data”.

Test A/B, es una técnica en la que comparamos un grupo de control con una variedad de grupos de test para determinar qué cambios o tratamientos producirán una mejora dada una variable objetiva (por ejemplo, una ratio de respuesta de una acción de marketing). Un ejemplo de este experimento de testing A/B (también llamado split testing o bucket testing), es determinar qué texto, maquetación, imágenes o colores producen una mejora en las ratios de conversión de una tienda online o una acción de marketing por email. El big data nos permite ejecutar y

analizar una gran cantidad de pruebas, siempre asegurando que los grupos son de un tamaño suficiente para detectar diferencias estadísticamente significativas entre el grupo de control y los grupos de pruebas. Cuando manipulamos más de una variable en el experimento simultáneamente, la generalización multivariante de esta técnica, que se aplica a modelos estadísticos, se le llama A/B/N testing. (Ladredo, 2018)

Las **reglas de asociación** son un conjunto de técnicas que permiten descubrir relaciones interesantes. En minería de datos y aprendizaje automático, las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos, por ejemplo, entre variables de varias bases de datos enormes. Estas técnicas consisten en aplicar una variedad de algoritmos para generar y testear las pautas posibles. Una aplicación práctica sería el análisis de la cesta de la compra de un comerciante online, en la que podemos determinar qué productos son comprados conjuntamente con frecuencia, para realizar acciones de marketing eficientes. Por ejemplo, a priori quizá no se nos hubiera ocurrido, pero se ha descubierto que un producto que se compra en los supermercados junto con los pañales es la cerveza.

Análisis cluster (o de conglomerados) es un método estadístico para clasificar objetos separando un grupo diverso en grupos más pequeños de objetos similares, cuyas características de similitud son conocidas previamente. Un ejemplo de análisis cluster ayuda a segmentar a los consumidores en grupos similares para realizar acciones de marketing segmentadas.

Crowdsourcing que se podría traducir como «colaboración abierta distribuida» o «externalización abierta de tareas», se trata de una técnica de recogida de datos facilitada por una comunidad o gran conjunto de gente conectada en torno a la red que llevan a cabo una tarea conjunta.

Fusión e integración de datos. (Ladredo, et, al; 2018) . Son una serie de técnicas que permiten integrar y analizar datos de múltiples fuentes con el objeto de realizar descubrimientos entre la información de manera más eficiente y potencialmente más precisa que si fueran analizados utilizando una sola fuente de datos. Un ejemplo práctico sería la aplicación combinada de diversos sensores de datos de

dispositivos conectados en la llamada Internet de las cosas, integrado con el rendimiento de sistemas complejos distribuidos en una explotación petrolífera. Otro ejemplo sería el análisis via procesamiento de lenguaje natural de datos de redes sociales combinados con datos de ventas en tiempo real, con el objetivo de determinar el efecto que está teniendo una campaña de marketing en el sentimiento de los clientes y su comportamiento reflejado en las decisiones de compra.

Data mining. Consiste en extraer patrones de grandes datasets mediante la combinación de métodos estadísticos y de aprendizaje automático con la gestión de las bases de datos. Entre las técnicas de datamining se incluyen técnicas de aprendizaje de reglas de asociación, análisis de agrupamiento, clasificación y regresión. Como ejemplos de aplicaciones prácticas estarían la minería de datos de clientes para determinar qué segmentos son más proclives a responder a una oferta, minar datos de recursos humanos para identificar características de los empleados de más éxito, o el análisis de cestas de compras para modelar el comportamiento de compras de los clientes. (Ladredo, et, al; 2018)

Aprendizaje mediante ensembles (ensemble learning). Consisten en utilizar múltiples modelos predictivos, ya hayan sido desarrollados mediante estadística o aprendizaje automático, para obtener mejores predicciones de rendimiento que puedan ser obtenidos de cualquiera de los modelos constitutivos. Son un tipo de aprendizaje supervisado.

Algoritmos genéticos. Es una técnica utilizada para optimizar datos inspirada en el proceso de la evolución natural o supervivencia de los mejor adaptados. Con esta técnica las soluciones posibles son codificadas como si fueran cromosomas que pueden combinarse y mutar. Estos cromosomas son seleccionados y separados para sobrevivir dentro de un ecosistema modelado que determina la adaptabilidad o el rendimiento de cada uno dentro del conjunto. Estos algoritmos evolutivos funcionan bien para solucionar problemas no lineales, como, por ejemplo, mejorar la planificación de tareas en la industria manufacturera, o la optimización del rendimiento de una cartera de inversión.

Redes neuronales. Los modelos computacionales, inspirados por los trabajos de redes neuronales biológicas, como las conexiones de las células del cerebro, que

buscan patrones entre datos. Las redes neuronales son apropiadas para buscar patrones no lineales y optimización. Entre las aplicaciones prácticas de esta técnica, por ejemplo, la identificación de los clientes de alto valor que están en riesgo de cambiar de proveedor, o la identificación de partes de seguro fraudulentos.

Análisis de redes. (Ladredo, et, al; 2018). Son técnicas empleadas para caracterizar relaciones entre nodos separados en un gráfico o red. Al analizar las conexiones entre individuos de una comunidad en las redes sociales podemos extraer cómo fluye la información o quién ejerce la mayor influencia y sobre quiénes. Entre las aplicaciones prácticas están la identificación de los líderes de opinión para realizar una acción de marketing precisa, o identificar los cuellos de botella en los flujos de información de las compañías.

2.3 Data Science.

Con el auge de los grandes datos surge un nuevo concepto, ciencia de datos, que se utiliza para referirse a una gama de técnicas necesarias para procesar y manipular grandes cantidades de información de las estadísticas y la informática. También se incluye el surgimiento de un nuevo perfil profesional, "Data Scientist", donde quienes se formen en esta profesión deberán comprender los negocios, las herramientas computacionales y el análisis e interpretación estadística. (Hernández-Leal, E. J.; 2017).

(Wil M.P. van der Aalst, 2014) revela que el principal objetivo de la ciencia de datos es responder a las cuestiones “¿Qué pasó?", "¿Por qué sucedió?", "¿Qué pasará?" y "¿Qué es lo mejor que puede pasar?".

Además de habilitar nuevos modelos comerciales, la ciencia de datos se puede utilizar para hacer las cosas de manera más eficiente o rápida.

Optimizar el recorrido del cliente es una de las muchas formas en que las organizaciones se benefician de la ciencia de datos y extraen valor de los datos. El aumento de la competencia hace que la ciencia de datos sea un diferenciador clave. Las organizaciones que no utilicen los datos de forma inteligente no sobrevivirán.

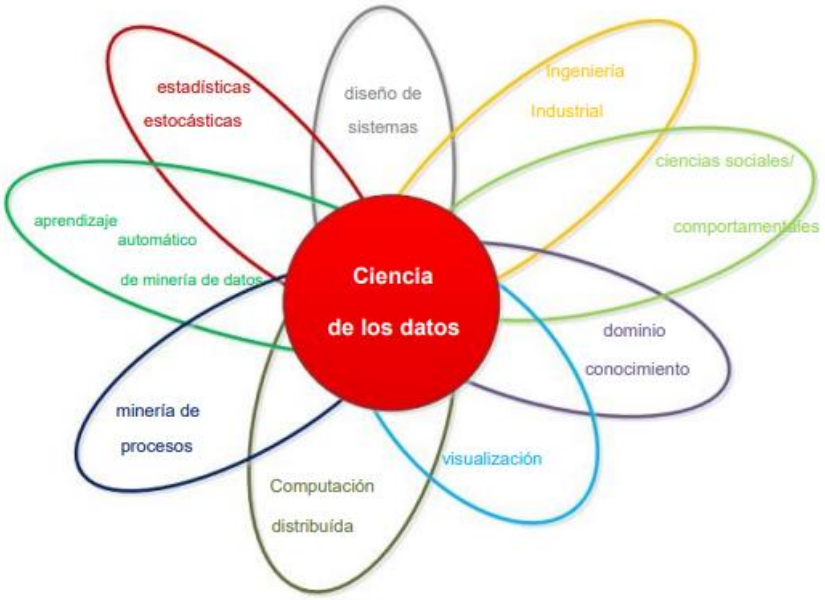


Fig. 2. Perfil del científico de datos: se combinan diferentes subdisciplinas para crear un ingeniero con habilidades cuantitativas y técnicas, creativo y comunicativo, y capaz de realizar soluciones integrales.

2.4 Datamining

Análisis y gestión de procesos orientados a la experiencia del cliente

Cada vez más se observan los esfuerzos orientados a adecuar las organizaciones al complejo escenario en que se mueven. Cambios de reglas de juego, incremento de la competencia, apertura al mundo a través de la tecnología, hacen al cliente mucho más exigente, modificando sus demandas y necesidades. La Gestión basada en los Procesos, surge como un enfoque que centra la atención sobre las actividades de la organización, para optimizarlas. En este trabajo se considerará a la organización como una red de procesos interrelacionados o interconectados, donde la estructura organizativa vertical clásica, eficiente a nivel de funciones, se orienta hacia una concepción horizontal, desplazándose el centro de interés desde las estructuras hacia los procesos, como metodología para mejorar el rendimiento, concentrándose en el diseño disciplinado y cuidadosa ejecución de todos los procesos de una organización. Concluyendo, la metodología de aplicación de la Gestión basada en Procesos se trata de una herramienta de gestión adecuada para el momento actual, constituyéndose con fuerza como una alternativa exitosa para la obtención de resultados cada vez mejores.

La experiencia del cliente (CX por sus siglas en inglés) se ha convertido en el foco de atención de todas las marcas en los últimos años, ya que las organizaciones han notado la importancia que tiene esto para mantener a los consumidores satisfechos y leales a la marca.

La experiencia del cliente es el conjunto de experiencias que tiene un cliente con una marca durante el tiempo en el que haya una relación de servicio. Mientras mejor sea la experiencia del cliente, mayor es la probabilidad de incrementar la retención de los clientes, generar nuevos leads, incrementar el ticket promedio, etc.

Bajo este enfoque, la estructura organizativa vertical clásica, eficiente a nivel de Funciones, se orienta hacia estructuras de tipo horizontal, tal cual lo define Ostroff (2000) quien sostiene que no hay contraposición entre modelos, y que cada empresa debe buscar su equilibrio en función de sus propias necesidades y posibilidades.

Revisión de la literatura

Empresas y organizaciones están experimentando un proceso que Vandermerwe y Rada (1988) denominan "servitización", una estrategia basada en la diferenciación en servicios. Romero Amado (2014) apunta a que esta situación obedece a un proceso de desindustrialización, debido al crecimiento económico, la innovación tecnológica y la demanda de servicios en los que el acceso al beneficio es más importante que la propiedad del producto.

Esto supone una estrategia que emplean las empresas para competir en el mercado, y asegurar su sostenibilidad y éxito financiero a largo plazo, la cual está basada en la diferenciación en servicios como oferta central y donde los clientes ven a la organización como proveedora de servicios (Gebauer, Gustafsson, & Witell, 2011, p. 1272). Es decir, en una estrategia que se desarrolla como ventaja competitiva para mejorar el desempeño organizacional (Bozkurt & Kemer, 2014), que, de acuerdo a Gannon, Lynch y Harrington (2014), implica una ejecución impecable para crear o hallar la ventaja -si existe o es posible conseguirla-.

La diferenciación de servicios puede ser contemplada como fuente de ventajas competitivas cuando se considera un proceso donde se combinan personas, ideas habilidades y características únicas entendidas como recursos especiales y de difícil imitación (Bharadwaj, Varadarajan, & Fahy,1993, p. 83; Oliveira Teixeira & Werther Jr., 2013). Así entonces, la fuente de ventaja competitiva es multifactorial y puede venir de cualquier combinación de elementos difícilmente imitables, e ideas que requieren inversión y sinergia por parte de la organización para ser desarrolladas (Dierickx & Cool,1989; Zaridis, 2009).

Modelos de estructuras organizacionales para la gestión del servicio

La estructura organizacional de cada empresa es única por la cultura, el mercado en el que compite y la visión de su directiva (Csaszar, 2012). Sin embargo, a partir de prácticas de prestación de servicios se pueden rastrear elementos estructurales y encontrar puntos de convergencia que permitan caracterizar modelos de servicios.

División entre creación de productos y prestación de servicios

Según Gebauer, Pütz, Fischer y Fleisch (2009, p. 109), el crecimiento en servicios parece requerir una organización separada y distinguible, porque el negocio del servicio debe controlar completamente la identificación y desarrollo de los clientes, el precio y la entrega de las ofertas. Es decir, las organizaciones están concentradas en una función productiva o manufacturera y añaden ofertas de servicios complementarias, no la inclusión del área de oferta central, lo cual sucede por no entender particularidades de la entrega de servicios y, más bien, centrarse en atender quejas, reclamos o garantías que no generan ingresos.

Gestión por procesos

Para realizar mediciones internas una alternativa completa es la implementación del CRM (Customer Relationship Management) que más que un software es la creación de una cultura corporativa que trabaja poniendo al cliente como prioridad, para conocerlo, entenderlo, atenderlo y satisfacerlo, a través de un equilibrio entre, los

procesos, la tecnología, el recurso humano, y el uso eficaz y eficiente de los demás recursos asignados (Universidad Militar Nueva Granada, 2010, pág. 53).

El resultado final de esta estrategia CRM debe ser adquirir nuevos clientes, fidelizar a los clientes a través de la atención oportuna de sus necesidades, afianzar la relación con los clientes actuales al exceder sus expectativas, retener a aquellos clientes que por errores en la cadena de proceso o por factores externos desean retirarse de la compañía.

El CRM debería permitir a la organización conocer aspectos de su cliente que le ofrezcan información demográfica, económica y emocional, segmentar su base de clientes para mayor precisión al momento de ofertar nuevos productos, de fidelizar y de retener, conocer las insatisfacciones de sus clientes para atacar la causa raíz de sus problemas con la organización y crear un modelo único de interacción que convierta a sus clientes en algo más que clientes.

Enfoque Basado En Procesos

Cuando se habla de eficacia hacemos referencia al logro de los resultados que se esperaban tras el establecimiento de los objetivos de la empresa, por otra parte, la eficiencia se orienta al uso adecuado y la optimización de los recursos para el cumplimiento de dichos resultados (Universidad Militar Nueva Granada, 2013, pág. 35); si dentro de la organización no se logra que los procesos se realicen bajo estos dos pilares, podemos decir que se está conduciendo de manera inadecuada el rumbo de la organización.

¿Por qué son importantes los procesos? Si evaluamos el funcionamiento de una empresa, son múltiples las actividades que deben realizarse para la entrega del producto o servicio al consumidor final y más aún para lograr en él una experiencia de satisfacción total, si bien las empresas en el pasado se preocupaban solo por vender un producto terminando con estándares básicos de calidad, los aspectos de competencia, costos, clientes entre otros de la actualidad no permiten a las empresas de hoy conformarse con un producto de baja calidad, ni mucho menos a los clientes aceptarlo. (James F. Riley, 2001, pág. 6.1)

Mejora Continua

La mejora continua, bajo el concepto definido en la norma ISO 9000 tiene por objetivo aumentar la satisfacción de los clientes y de otras partes interesadas, la mejora debe ser una actividad continua al interior de la organización y deriva de la información obtenida de clientes, auditorias, revisión del SGC, que suministran el insumo de todo aquello que tiene una oportunidad de ser mejorado para el bien del cliente y de la empresa. (Organización Internacional de Normalización, 2005, pág. 6)

Elementos del proceso

Los elementos que conforman un proceso son:

1. Inputs: recursos a transformar, materiales a procesar, personas a formar, informaciones a procesar, conocimientos a elaborar y sistematizar, etc.

2. Recursos o factores que transforman: actúan sobre los inputs a transformar.

a) Factores dispositivos humanos: planifican, organizan, dirigen y controlan las operaciones. La Gestión por Procesos: Un Enfoque de Gestión Eficiente

b) Factores de apoyo: infraestructura tecnológica como hardware, programas de software, computadoras, etc.

3. Flujo real de procesamiento o transformación: La transformación puede ser física (mecanizado, montaje etc.), de lugar (el output del transportista, el del correo, etc.), pero también puede modificarse una estructura jurídica de propiedad (en una transacción, escrituración, etc.) (Ilustración 1)

2.5 Big Data Analytics

En Big Data existen tres términos conocidos como las 3 V: Volumen, Velocidad y Variedad.

No solo hace referencia a los problemas relacionados con los datos, sino que también incluye un amplio espectro de técnicas, tecnologías, métodos y paradigmas no convencionales que apoyan la solución de problemas relacionados con datos de una forma diferente y, generalmente, más adecuada que los métodos tradicionales.

Permitió entonces nuevas y mejores formas de procesar la información, con ventajas sobre los enfoques tradicionales, los cuales no responden de forma adecuada sobre las necesidades actuales de las compañías en términos de velocidad, costos de implementación, escalabilidad, flexibilidad y elasticidad sobre entornos más complejos.

Orientados principalmente a la computación distribuida y el procesamiento paralelo masivo, que han convergido de cierta forma con tecnologías como la computación en la nube (“Cloud Computing”) y las nuevas formas de almacenar los datos, mediante modelos no relacionales, sobre todo cuando se deben tener en cuenta los costos de dicho almacenamiento para su posterior procesamiento.

Big Data adecuadamente en una organización (privada, gubernamental, o de cualquier sector) pudiera representar una ganancia, lo verdaderamente importante para esta organización es el Valor que se puede generar partir de estos datos, siendo esto aún más importante, si se parte de la premisa que indica que, en la mayoría de los casos, los datos no siempre generan este valor esperado, por lo que las organizaciones deben estar en la capacidad de descubrir esto en tiempo real sin descuidar aspectos de seguridad, integración, funcionalidad y otros atributos de calidad que sean contemplados en cada dominio en particular.

“Big Data Analytics”

Transformar la Big Data en conocimiento y llevar estas aplicaciones a las organizaciones, las cuales agregan retos adicionales como lo son: El costo computacional, la seguridad informática, la integración con otros sistemas, la volatilidad de los requisitos y demás aspectos propios de cada negocio o área de dominio.

Una definición de cuatro conceptos clave que giran en torno a Big Data Analytics, la arquitectura general para analizar Big Data, adoptada por la mayoría de las organizaciones estudiadas y las oportunidades, retos y tendencias que se encuentran en curso, en relación con Big Data Analytics

“explosión de datos” conocida generalmente como “Big Data, conlleva a nuevos desafíos (como la seguridad informática y la minimización de costos de almacenamiento y procesamiento) que obligan a dar una mirada más extensa y

global a las alternativas que se proponen y, sobre todo a las necesidades puntuales de cada organización o unidad productiva. III. BIG DATA A

Big Data no solo hace referencia a los datos si no a las técnicas, métodos y tecnologías utilizadas para solucionar un problema, lo cual aleja un poco esta definición de las 3 Vs.

Oportunidad para entender el mundo desde los diferentes dominios o áreas de aplicación en donde se generan datos que pueden ser transformados en información útil para la toma de decisiones.

APLICACIONES DE BIG DATA ANALYTICS

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

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

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

D. Gobierno y sector público El sector público también involucra problemas de Big Data en la medida en que las poblaciones de las regiones (países, ciudades, comunidades) pueden ser grandes y con necesidades diversas o sectorizadas, en

donde cada individuo genera millones de datos a través de diferentes canales, generalmente públicos (como los servicios públicos y las redes sociales, por ejemplo). Con Big Data Analytics es posible obtener resultados en tiempo real que permitan establecer políticas públicas que respondan a las diversas necesidades de manera efectiva y eficaz.

E. Sector salud Es uno de los sectores más importantes para el desarrollo de las regiones, pero también es uno de los más complejos en términos de datos, ya que contempla información clínica, la cual, en un futuro podrá contener más datos de los que hoy se contemplan (por ejemplo, información genética); también contempla datos farmacéuticos, datos de prácticas, preferencias y hasta registros financieros de los pacientes. La integración de todos estos datos será un factor clave para tener un sistema de salud mejor y más asequible (sin entrar en discusiones políticas). Con Big Data Analytics podrá ser posible tener una vista 360 grados no solo de pacientes sino también de organizaciones de salud, así como también se podrá optimizar el funcionamiento de las entidades que componen este sistema (por ejemplo, hospitales e intermediarios).

F. Servicios públicos y Telecomunicaciones En la prestación de servicios públicos como el agua, gas, electricidad y telecomunicaciones, puede usarse Big Data Analytics como la base para la detección de filtraciones, fraudes y pérdidas no-técnicas, así como para realizar “mediciones inteligentes”, partiendo de que los datos que son generados por los dispositivos y sistemas involucrados en la prestación de estos servicios son caracterizados por encontrarse en ambientes de Big Data. G. Sector manufacturero En este sector se pueden utilizar los datos procedentes de las máquinas de manufactura (generados por sensores, por ejemplo) para que, combinados con el efectivo análisis de la demanda, permitan una producción óptima, minimizando el desperdicio y el re-trabajo. Esto es posible a través de métodos relacionados con Big Data Analytics, que proporciona técnicas de avanzada para apoyar la toma de decisiones sobre estos datos caracterizados por ser complejos.

Preguntas de la unidad II

1.-Busca datos y terner un obvjetivo definido es una funcion de:

2.-Qué son las técnicas de recolección de datos: EXECPO

3.-Es su principal funcion alanizar y filtrar datos y diseña estrategias en una empresa.

5 Proceso quer implementa siertos funcionamientos principales para un buen desarrollo.

5.-¿Cómo hacer un análisis de Big Data eficaz?

6.-principal funcion de una data split

7.-una empresa Busca datos y tiene un obvjetivo definido es una funcion de?

8.-Consiste en comparar una serie de acciones y observar las reacciones de los usuarios ante un determinado producto o mensaje

9.-los algoritmos identificar patrones complejos entre gran cantidad de datos, infiriendo así sus propias reglas para detectar patrones similares en nuevos conjuntos de datos

10.-La extracción de información útil de grandes volúmenes de conjuntos de datos. Al contrario, la ciencia de datos utiliza algoritmos de Machine Learning

1.-R=big data, 2.- R=Dejar datos abiertos, 3.- R= Data Analyst 4.- R= Datos y estrategias definidas 5.-R=Producto y producción 6.-R=Fragmentos en celdas, 7.-R=proceso de una empresa, 8.-R=Proceso de Análisis de datos, 9.-Ordenados en columnas y filas 10.- ML,DL

REFERENCIAS

Aceves, J, García, S. y González, M. (2011) Revista Neurología, Neurocirugía y Psiquiatría. 2011; 44(4): Sep-Dic: 128-132

Aguilar, L. J. (2013). Big Data, Análisis de grandes volúmenes de datos en organizaciones. Alfaomega Grupo Editor.

Big Data Analytics Transforming Data Into Value Actian." [Online]. Available: http://www.actian.com/. [Accessed: 11-May2014].

Buzzoni, M. (2010) Rethinking Popper and His Legacy. International Studies in the Philosophy of Science Vol. 24, No. 3, September 2010, pp. 309–321

Cerda J. y Villarroel L. (2007) Interpretación del test de Chi-cuadrado ($\chi 2$) en investigación pediátrica Rev Chil Pediatr 2007; 78 (4): 414-417

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

De la Cuesta, C. (2006) Teoría y método la teoría fundamentada como herramienta de análisis. Cultura de los Cuidados. Semestre 2006 • Año X – N.° 20

Díaz, C. (2007) Entrevista a la Doctora en Educación Carmen Díaz Bazo

Charmaz, K., (2005). Grounded theory in the 21st Century. En: The Sage handbook of qualitative reserach (Denzin N K y Lincoln Y S). SAGE, Thousand Oaks, CA, pp.507-535.

Corbetta, P. (2007) Metodología y técnicas de investigación social.

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

"Big Data - Expertos EN IIC." *Instituto De Ingeniería Del Conocimiento*, 14 July 2021, https://www.iic.uam.es/big-data/.

Ellen A. Skinner1 and Melanie J. Zimmer-Gembeck 2. (2007). The Development of Coping. 24-03-2022, de Anonimo Sitio web: https://www.annualreviews.org/doi/pdf/10.1146/annurev.psych.58.110405.085705

Escamilla-Quinta, M. et. al. (2008) El cinismo: una estrategia de afrontamiento diferencial en función del género. *Psicothema 2008. Vol. 20, n'4, pp. 596-602*

Fernández, G. (2007) Metodología de la investigación. Universidad de Londres

Hernández, R., Fernández, C. y Baptista, P. (2010) Metodología de la Investigación. 5ª ed. México: Mc Graw-Hill.

Kaushik, Avinash.(2011). Analítica Web 2.0, Barcelona: Gestión 2000.

Ladrero, I. (2018, 15 noviembre). *Técnicas de análisis Big Data*. BAOSS. Recuperado 17 de marzo de 2022, de https://acortar.link/Ua1RcM

Lovett, J. (2011). Social Media. Métricas y análisis, Madrid: Anaya.

Ortiz, F. (2010) Metodología de la investigación: el proceso y sus técnicas. México: Limusa Noriega.

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

Ramírez, R. (2010) Proyecto de investigación: como se hace una tesis**.** Lima: Academia de Magisters del Perú.

Rodríguez, et. al. (2010) Desarrollo y validación de una escala para medir imagen corporal en mujeres jóvenes. Salud Mental 2010; 33:325-332

Román, S. y Rodríguez, R. (2011) ¿Cómo podemos distinguir a los vendedores éticos de los que no lo son?: Implicaciones para el proceso de selección y formación de los comerciales. Cuadernos de Gestión Vol. 11. Especial Responsabilidad Social (Año 2011), pp. 85-99

Secretaría de Marina Armada de México (2010) Manual para elaborar y evaluar trabajos de investigación

Stecher, A. (2010) El análisis crítico del discurso como herramienta de investigación psicosocial del mundo del trabajo. Discusiones desde América Latina. Bogotá, volumen 9 número 1, enero abril 2010, pp 93-107

Strauss, A. (2004) Anselm Strauss en conversación con Heiner Legewie y Barbara Schervier-Legewie. Forum Qualitative Social Research,(on line journal) 5 (3) Art. 22. Disponible en: http://www.qualitative-research.net/fqstexte/3-04/04-3-22b-s.htm

Triola, M. (2009) Estadística. 10 edición. Editorial Pearson

Van Dijk, T. (2003a). Prólogo. En L. Iñiguez (Ed.), *Análisis del Discurso. Manual para las Ciencias Sociales* (pp. 11-16). Barcelona: UOC

Van Dijk, T. (2003b). La multidisciplinaridad del Análisis Crítico del Discurso: un alegato a favor de *Métodos de análisis crítico del discurso* pp. 143-17

Westreicher, G. (2020, 23 septiembre). *Análisis de datos*. Economipedia. Recuperado 16 de marzo de 2022, de https://acortar.link/1A4N3l

Acuña Bermeo, C. F. (2008). *Diseño, propuesta de mejora y estandarización de los procesos orientados al cliente de la empresa Edinun* (Bachelor's thesis, QUITO/EPN/2008).

Bernal Zipa, M. M. (2015). Gestión por procesos y mejora continua, puntos clave para la satisfacción del cliente.

Martínez, L., & El Kadi, O. (2019). Logística integral y calidad total, filosofía de gestión organizacional orientadas al cliente. *Revista Arbitrada Interdisciplinaria Koinonía*, *4*(7), 202-232.

Pérez Fernández, J. A. (1999). *Gestión de calidad orientada a los procesos* (No. HF5549 P415).

Unidad III. Desarrollo de Big Data
OBJETIVOS: Aplicar técnicas de análisis de datos a gran escala; para la toma de mejores decisiones empresariales, la generación de oportunidades de negocio y la optimización de recursos.

3.1 Diseño de investigación

Para empezar, se muestran definiciones sobre el concepto de Diseño de investigaciones:

Según Diseño de investigación Elementos y características (2018) el diseño de investigación se define como la elección de métodos y técnicas por parte del investigador para combinarlos de manera lógica y lógica con el fin de abordar de manera efectiva la pregunta de investigación. El diseño es una guía sobre "cómo" realizar una investigación utilizando un método particular. Cada investigador tiene una lista de preguntas para evaluar. Un diseño de investigación se puede utilizar para preparar un esquema de cómo se llevará a cabo la investigación. Por lo tanto, la investigación de mercado se realizará sobre la base del diseño de investigación.

Características del diseño de investigación

El diseño del estudio tiene 4 características clave:

Neutralidad: Los resultados previstos en el diseño deben ser imparciales y neutrales. Conocer las opiniones y conclusiones de múltiples personas sobre la nota de la evaluación final, y considerar aquellas que estén de acuerdo con los resultados obtenidos.

Confiabilidad: si el estudio se lleva a cabo regularmente, se espera que los investigadores involucrados calculen resultados similares cada vez. El diseño del estudio debe mostrar cómo se formuló la pregunta de investigación para garantizar los criterios de resultado obtenidos, y esto solo sucederá si el diseño del estudio es sólido.

Validez: Existen múltiples herramientas de medición disponibles para el diseño, pero las herramientas de medición válidas son aquellas que ayudan al investigador a medir los resultados de acuerdo con el objetivo de la investigación y nada más. El

cuestionario desarrollado a partir de este diseño de investigación será entonces válido.

Generalización: El resultado del diseño debe ser aplicable a una población y no sólo a una muestra restringida. La generalización es una de las características clave del diseño de la investigación.

Según (¿Qué es el Diseño de Investigación y cómo se realiza?, 2019). Hay 3 tipos de diseños de investigación. Los describiremos en detalle a continuación.

1. Diseño Experimental

Los diseños de investigación experimental están diseñados para tener el mayor grado de control (por parte del investigador).

2. Diseño de contraste

El diseño comparativo, a su vez, se divide en dos: correlación (cuando existe algún grado de relación entre las variables; no admite causalidad) y el diseño comparativo propiamente dicho (en el que la variable independiente es una elección; es decir, el tema con su valor "post" [por ejemplo, raza o género]). Por otro lado, con un diseño comparativo apropiado, se puede establecer una relación cuasi-causal.

3. Observación/Diseño de Encuesta

Este tipo de diseño de investigación tiene la menor cantidad de control por parte del investigador, es decir, no hay manipulación, solo se observa.

Un ejemplo de un diseño de investigación observacional es una encuesta.

Las variables en la investigación

Otro concepto importante que debemos conocer para entender bien qué es el diseño de investigación son las variables en investigación, ya que todos ellos las tienen.

1. Variables dependientes

La variable dependiente, que se suele expresar mediante “Y”, es el efecto que se produce a partir de la variable independiente. Por ejemplo, puede ser el grado de ansiedad (que aumenta o disminuye en función de un tratamiento).

2. Variables independientes

Las variables independientes, sin embargo, se representan mediante “X”, y son la causa de los efectos. Es decir, siguiendo el ejemplo anterior, se trataría de los tratamientos psicológicos (variable independiente), por ejemplo, que influyen en el grado de ansiedad (variable dependiente). (s/f)

DISEÑO METODOLÓGICO

Conjunto de procedimientos para dar respuesta a la pregunta de investigación y comprobar la hipótesis. Plan o estrategia concebida para dar respuesta al problema y alcanzar los objetivos de investigación (Christensen citado por Bernal, 2000).

El diseño está determinado por el tipo de investigación que se va a realizar (Bernal, 2000). Estructura u organización esquematizada que adopta el investigador para relacionar y controlar las variables de estudio (Sánchez Carlessi, 1990).

Fig. 1. Tipos de Diseños de Investigación

EXPLORATORIO

- Para aproximarse a temas poco estudiados y sentar las bases para futuras investigaciones.

DESCRIPTIVO

- Su propósito es identificar las características, propiedades, dimensiones y regularidades del fenómeno en estudio.

EXPLICATIVO

- Su propósito es investigar por qué ocurren y en qué condiciones se manifiestan los fenómenos físicos y sociales.

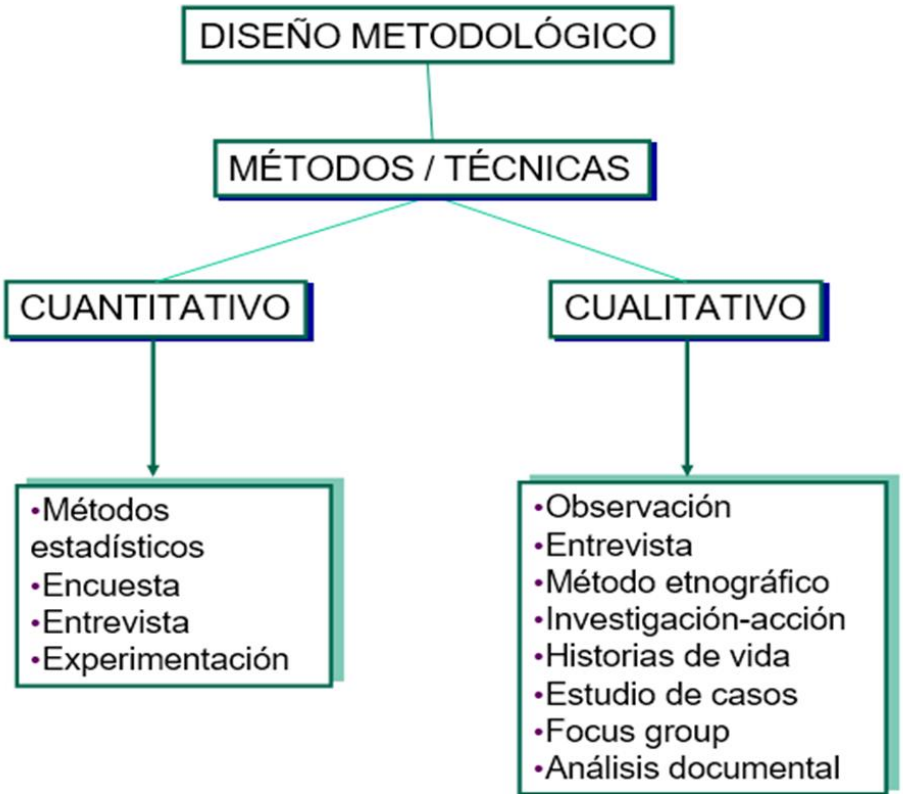


Fig. 2. Diseños metodológicos

Meta-análisis

Finalmente, otro concepto interesante de conocer en relación con los diseños de investigación es el meta-análisis; se trata de una técnica para evaluar los resultados cuantitativos de un conjunto de estudios empíricos.

Se trata de un tipo de metodología consistente en la revisión sistemática de diferentes estudios y resultados; a través de ella, se procede a aplicar una serie de técnicas estadísticas con objetivo de analizar dichos resultados, principalmente de forma cuantitativa. (Diseño de Investigación, 2019).

3.2 Aplicaciones de negocio Big Data

Según Joyanes L (2013) menciona como ejemplo a Apache Hadoop es una biblioteca de software de código abierto (open source) que soporta el procesamiento distribuido de grandes conjuntos de datos a través de miles de computadoras ordinarias. El proyecto Apache Hadoop ha nacido de la mano de las dos grandes empresas de la Web, Google y Yahoo!, cuyos investigadores trabajaron con grandes volúmenes de datos en grandes clusters de computadora.

Hadoop es el líder en plataformas de Big Data y su uso crece de modo espectacular, por no decir exponencial. Hadoop consta de tres componentes principales: Hadoop Distributed File System (HDFS), MapReduce y Hadoop Common. Además, existen otras tecnologías complementarias como HBase, Hive, Pig, y otras con la misma filosofía tales como IMPALA de Cloudera, DRILL o Google Big Query.

La integración de los datos dentro de las organizaciones, y sobre todo la integración con las herramientas de analítica de negocios, se construyen generalmente sobre fuentes de datos de archivos estructurados y relacionales que no aprovechan las ventajas de la arquitectura de datos masivamente escalables de Hadoop, plataforma fundamental de Big Data. Esta es una de las razones por las cuales las organizaciones a veces no afrontan con decisión el paso a estrategias de Big Data. Por este motivo, se requerirán herramientas que aprovechen toda la potencialidad de Hadoop para que la integración de datos se produzca de la manera más rápida, eficaz y lo más económica posible.

El movimiento de datos a granel se realiza con tecnologías tales como ETL (extraer, transformar, cargar) que extraen datos de una o más fuentes de datos (normalmente EDW, data ware houses de empresa), cargan los datos en una base de datos destino y los transforman en la base de datos destino. Las herramientas ETL se utilizan con frecuencia con Hadoop para aprovechar su potencia de procesamiento paralelo masivo.

Las aplicaciones del big data, han variado conforme a los años, pues con el paso del tiempo se ha podido tener acceso a más datos, muchísima más información

para analizar, y algunas de las aplicaciones que se tienen, varían según los objetivos de su análisis y los autores, por ejemplo;

Según Pérez M (2015), existe una gran variedad de aplicaciones de las técnicas de Big Data, pues siempre que sea necesario extraer el conocimiento inmerso en grandes volúmenes de datos, tienen cavidad las aplicaciones de big data. Este autor nos muestra opciones como:

- Patrones de detección de fraude
- Patrones de social media
- Patrones de modelado y gestión de riesgo
- Sector de la energía
- Big data en el Call Center

Además, Orteiza D (2019), menciona que las aplicaciones principales se desarrollan en los siguientes ámbitos;

1. Analítica para negocios, centrada en enfoques descriptivos, predictivos a predictivos, típicamente apoyando el descubrimiento y presentación de patrones relevantes en problemas con grandes conjuntos de datos.
2. Analítica para políticas públicas, pues en base al análisis de casos hospitalarios, propiedades de los datos, cámaras de vigilancia o información de posicionamiento en tiempo real, se puede realizar una toma de decisiones para la elaboración de políticas públicas, que puedan optimizar el uso de recursos

La aplicación de las tecnologías Big Data en empresas aporta esencialmente una capa de inteligencia al negocio. En muchos casos se adoptan para una gestión eficiente de los recursos (materiales o humanos), gracias a un análisis detallado de grandes volúmenes de datos de la empresa como facturación, clientes, productos, fechas, etc. El término Big Data continúa siendo el vocablo de moda en entornos empresariales y mediáticos, en cualquier momento los aspectos más cotidianos de nuestras vidas se ven afectados por el Big data. Dentro de esto hay unas áreas donde el Big data está marcando la diferencia como lo es:

- **Entendiendo y segmentando a los clientes**

Donde se tiene que Marketing y ventas son quizá las áreas de mayor aplicación de Big data en la actualidad. Los datos se utilizan para comprender mejor a los clientes, sus comportamientos y preferencias. Las empresas están dispuestas a ampliar los centros de datos tradicionales con los de redes sociales, logs de navegación, análisis de textos y datos de sensores para obtener una imagen completa de su cliente. El objetivo principal es en la mayoría de los casos es crear modelos predictivos. Los hipermercados pueden predecir mejor qué productos se venderán mejor, y las aseguradoras de coches pueden comprender mejor cómo conducen sus clientes. Incluso las campañas electorales pueden optimizarse gracias a Big data analytics.

- **Entendiendo y optimizando los procesos de negocio.**

El Big data se está utilizando cada vez más para optimizar los procesos de negocio en las empresas. Los negocios están optimizando su stock basándose en predicciones generadas gracias a datos de redes sociales, tendencias de búsquedas en la web y predicciones meteorológicas. Un proceso que se está transformando particularmente gracias al Big data es el de la cadena de suministro y la optimización de rutas de reparto. Gracias al posicionamiento geográfico y sensores de identificación por radiofrecuencia se puede realizar un seguimiento de las mercancías y vehículos de reparto, optimizando las rutas, integrando datos de tráfico en tiempo real.

- **Trading financiero.**

Las actividades relacionadas con High-Frequency Trading (HFT) es donde se da el mayor uso del big data. Una serie de algoritmos para realizar decisiones de compraventa de valores por millones en fracciones de segundo, teniendo en cuenta además de las señales tradicionales que tienen en cuenta los traders humanos como análisis técnicos, comportamientos de materias primas, resultados de empresas, sectores, índices, … se le añaden noticias en tiempo real, mensajes de redes sociales, foros, declaraciones públicas de personalidades, etc. Es decir, un nuevo tipo de datos (estructurados y no estructurados) que con anterioridad al big data eran imposible de manejar. (Ladrero**,** 2020).

3.3 Métricas y objetivos específicos del Big Data en el negocio.

En esta era digital, nuevas tecnologías, conceptos y metodologías disruptivas se han ido introduciendo en el campo corporativo, cambiando completamente la forma en que operan las empresas. Un campo emergente que no debe dejar de mencionarse es el de Big Data ¿En qué medida los indicadores clave de rendimiento (también llamados KPIs) y el monitoreo del desempeño juegan un papel en la estructura corporativa actual?

En la era de Big Data, las organizaciones se enfrentan cada vez más a grandes desafíos para analizar datos estratégicos. Por lo tanto, deben definir los KPIs y agrupar los recursos para mantenerse competitivas. Debido a que todas las empresas quieren crecer, estas se deben enfocar en producir productos relevantes para el mercado y un marketing efectivo. Por lo tanto, se aconseja a todos definir exactamente cómo miden el logro de sus objetivos y en qué momento las actividades pueden considerarse un “éxito”. Para mí, la experiencia del cliente es un KPI que toda empresa debería tener en cuenta. Esto se ha convertido en un factor extremadamente importante para obtener éxito hoy en día, ya que una buena reputación es una de las formas más efectivas de marketing. Por otro lado, las empresas también deben ser capaces de cuidar y medir muchos otros factores de éxito. La escalabilidad juega un papel importante. Hoy en día, los KPIs se pueden encontrar en varios departamentos de una empresa: en producción, en marketing, recursos humanos y otras áreas estratégicamente importantes. Por lo tanto, la supervisión del desempeño se utiliza para el éxito y el control de las funciones, así como para garantizar la calidad.

(Stephanie Ospina. (2019))

¿Cuáles son los componentes clave a tener en cuenta al establecer objetivos para que los KPIs sean efectivos?

El error de aplicación más común que cometemos los profesionales del marketing es que consideramos todos los números y métricas que obtenemos a través de Google Analytics y otras herramientas como un KPI. Si bien, la mayoría de estas métricas son PIs (“Indicadores de rendimiento”), pueden no ser “Indicadores clave de rendimiento”. En este caso, el concepto “clave” es el que hace la diferencia. Solo deberíamos considerar los KPIs como tal si estos son la clave para el camino hacia el éxito, y ese es el caso si están dirigidos directamente a alcanzar la meta. Como resultado, muchas métricas no son indicadores clave de éxito. El acrónimo SMART se usa a menudo para identificar los KPIs y, en mi opinión, es una fórmula realmente útil. Esta define los KPIs como específicos, medibles, alcanzables, orientados a resultados y temporales. (Stephanie Ospina. (2019))

Si se desea configurar un marco de referencia de KPIs, se debe tener una visión integral de los datos y actividades y siempre preguntarse: ¿qué estamos haciendo para lograr el objetivo? ¿Cómo definimos el éxito? ¿Hay un punto de referencia? ¿Cómo medimos eso? Para encontrar estas respuestas, se deben combinar los objetivos principales, los objetivos secundarios, tácticas, medidas y actividades, así como las medidas variables, e intentar integrarlas en un marco analítico. Esta es sin duda una de las tareas más difíciles y desafiantes, ya que los KPIs a veces pueden ser muy complejos. Al formular los KPIs, siempre se debe tener en cuenta una segunda regla: cuanto mayor es la influencia de las actividades en el logro de la meta, más central es la medida subyacente para el monitoreo del desempeño. (Stephanie Ospina. (2019))

¿Cuáles son algunos aspectos específicos de una empresa que absolutamente recomiendas que se debieran rastrear/medir con los KPIs?

Desafortunadamente, no hay un conjunto de KPIs que sea transferible a todos los proyectos y empresas. Todas las empresas tienen que definir sus propios objetivos. En consecuencia, la identificación de los KPIs es una tarea específica y a veces difícil, y la implementación puede ser muy compleja e irregular. Sin embargo, una cosa es cierta: todas las empresas quieren ganar dinero al final del día. Por lo tanto, está claro y que los parámetros comerciales como el volumen de ventas, las ganancias, el ROI y el ROAS son siempre importantes. En mi opinión, uno de los KPIs más infravalorados sigue siendo la satisfacción del cliente, además de las dimensiones de tiempo. Todas las empresas prosperan gracias a sus clientes. Por lo tanto, la satisfacción del cliente siempre debe ser el objetivo de todas las actividades, independientemente si hablamos de ventas, marketing o gestión de productos. Creo en recomendar a las empresas a que inviertan mucho en la satisfacción de sus clientes y la midan. Los KPIs pueden ser el porcentaje de clientes satisfechos, el tiempo de respuesta del soporte y/o las ventas después de contactar a los clientes potenciales o las visitas a las páginas del centro de ayuda por unidad de tiempo. (Stephanie Ospina. (2019)) Para mejorar nuestro negocio gracias a los datos, hay tres conceptos clave que necesitamos aprender a diferenciar: KPI 's, métricas y objetivos. Nunky Nice. (2021).

El primer paso será fijarnos unos objetivos ¿qué quiero conseguir? (objetivos). Después debemos plantear ¿cómo conseguir los objetivos de mi negocio? (KPIS). Finalmente nos preguntaremos ¿cómo puedo saber mis resultados? (métricas). Nunky Nice. (2021).

Objetivos: Son las metas que quiero conseguir. Pueden ser objetivos estratégicos (indican de manera global las metas que se pretenden alcanzar a medio y largo plazo) u objetivos específicos (expresan de manera concreta y a nivel accionable metas que contribuyen al objetivo global).

KPI’s: Son una medida de valor aportado al negocio, normalmente resultado de la combinación de varias métricas. Un KPI es un indicador clave en el desempeño del negocio. Sirve para medir el éxito o el fracaso de las acciones que se llevan a cabo y están directamente relacionados con el cumplimiento de los objetivos de las compañías. Este indicador cambia si alcanzamos el objetivo o se modifica el foco de la gestión.

Métricas: Son los datos que miden una actividad. Una métrica es una medida registrada para evaluar aspectos y variables de una actividad o proceso para medir el éxito o el fracaso del desempeño en esa actividad concreta.

Un KPI es una métrica, pero una métrica no es necesariamente un KPI. Es importante diferenciar ambos términos puesto que los KPIs buscan potenciar el negocio al 100%, tienen que buscar un significado a todos los niveles de la organización y tienen que ser sencillos de entender (responder a una pregunta), tienen que provocar liderar una acción. Los KPIs son como un instrumento de navegación, para ayudar a los managers a saber si el viaje está siendo exitoso o nos separamos del camino próspero. Nunky Nice. (2021).

Un CSF es un factor crítico o una actividad necesaria para asegurar el éxito de una empresa u organización. Es recomendable identificar primero los CSFs (Critical Success Factors) de los objetivos de negocio. Estos CSFs no cambiarán si no cambian los objetivos de negocio. La dificultad está en identificar unos puntos vitales para nuestra empresa y momento, en lugar de coleccionar y reportar un número grande de datos que no aportan valor. Sólo hay que usar los que son relevantes a nuestra estrategia y compararlos con los de nuestra industria. Nunky Nice. (2021).

3.4 Los tableros de control y reportes Dashboard.

Tableros de Control.

Según Joyanes (2013). Los cuadros de mando o tableros de control (dashboards) son como los tableros o mandos de control de un automóvil y visualizan datos de un modo fácil de comprender. La información se presenta en gráficas, cartas y tablas que muestran el rendimiento real frente a métricas deseadas o informes de estado actual. Un cuadro de mando proporciona acceso fácil a información temporal (fecha y hora, timely) y acceso directo a la gestión de informes. Hoy día son muy populares. Algunas herramientas pueden ser: Microstrategy Dynamic Enterprise Dashboards (http://microstrategy.com/dashboards), Dashboard Bloomberg Terminal.

Concepto:

Según Fleitman (2015). El tablero de control es una metodología gerencial que sirve como herramienta para la planeación y administración estratégica de las empresas. Lo podemos definir como una estructura de control de la administración y operación general de la empresa, cuya fortaleza radica en su filosofía de mejora continua y en el trabajo en equipo basado en una visión estratégica unificada.

Implantación.

Al implantar el tablero de control se utilizan criterios de medición e indicadores para controlar la eficiencia y eficacia en el cumplimiento de la visión, misión y objetivos de la empresa.

Beneficios.

- Facilita la toma de decisiones a los socios y ejecutivos de una empresa ya que se tiene la información de manera inmediata de las diferentes áreas.
- Permite detectar inmediatamente las desviaciones de los planes, programas y estrategias y decidir las medidas correctivas.
- Mide el desempeño de la empresa en resultados financieros, atención, relación y satisfacción de los clientes, procesos internos, desarrollo y conocimiento.
- Facilita el control de los resultados financieros, midiendo simultáneamente el avance en el desarrollo de capacidades y la adquisición de activos intangibles relaciones con clientes, habilidades y motivación de los colaboradores.
- Pone énfasis en los indicadores financieros y no financieros y los incluye en el sistema de información para todos los niveles jerárquicos de la empresa.

Objetivos.

- Medir los avances y cumplimiento de la visión, la misión, los valores, los objetivos y las estrategias de la empresa.

- Integrar el plan estratégico de la empresa con los planes operativos de las áreas.
- Crear tableros de control para cada área y alinearlos con el tablero de control de la dirección.
- Desarrollar el tablero de control individual de cada puesto alineado con el tablero de control del nivel jerárquico inmediato superior.
- Identificar los diferentes tipos de indicadores existentes en un proceso (Indicadores de entrada, de salida, de eficiencia, de eficacia, de calidad, productividad, impacto y cultura).
- Alineamiento y realineamiento de la empresa a los cambios tecnológicos y de mercado. (Fleitman, et, al; 2015).

Para Orozco (2016). El tablero de control es una herramienta gerencial que tiene por objetivo principal presentar el estado actual de uno o varios elementos de la medición (indicadores, planes, estrategias, iniciativas) de la gestión de una compañía, bien sea a nivel global o por cada una de sus áreas o procesos.

Características.

A continuación, presentaremos algunas de las características cruciales para construir un buen tablero de control:

Visualmente claro: aunque de entrada parece algo trivial, una de las características más importantes de un buen tablero de control es su visualización.

Integral: imaginemos por un momento un comité de dirección en donde se tienen que presentar los avances en los indicadores financieros de la compañía, los proyectos de expansión y el estado del plan de marketing. (Ortiz, et, al; 2016).

Tipos.

Según el uso que se les puede dar en las organizaciones, los tableros de control se pueden clasificar en tres grupos:

- Operativo.

Su principal objetivo es llevar a cabo el seguimiento de los procesos o unidades de negocio de la organización, y de esta manera presentar y tomar decisiones a tiempo. Idealmente debe contener información, por ejemplo, del área financiera: ventas, cobros, cartera, producción.

- Directivo.

Este tipo de tablero tiene como propósito revisar los resultados internos de la organización por sus diferentes áreas, haciendo seguimiento a los indicadores de resultado y en una perspectiva a corto plazo. (Ortiz, et, al; 2016).

Dashboard.

Concepto.

Para Ortiz (2020). Un dashboard es una herramienta de gestión de la información que monitoriza, analiza y muestra de manera visual los indicadores clave de

desempeño (KPI), métricas y datos fundamentales para hacer un seguimiento del estado de una empresa, un departamento, una campaña o un proceso específico.

Características.

- Personalizado. Un dashboard debe contener únicamente los KPI que sean relevantes para el departamento, campaña o proceso que nos ocupa. Para orientarlo, podemos pensar en las preguntas principales a las que queremos responder. Por ejemplo, cuáles son las principales fuentes de tráfico a nuestra web, cómo está funcionando nuestro embudo de ventas o cuáles son los 5 productos que nos generan más ingresos.
- Visual. La idea de un dashboard es que podamos obtener la información que buscamos a golpe de vista. Por ello, los datos se presentan en forma de gráficos y debemos contar con indicadores rápidos a través de claves de color, flechas hacia arriba o abajo o cifras destacadas, por ejemplo.
- Práctico. La función principal de un dashboard siempre debe ser orientar las acciones de nuestro equipo. Por tanto, debe facilitarnos la información necesaria para que podamos saber cuáles son los siguientes pasos a seguir para mejorar los resultados.
- En tiempo real. A día de hoy, las acciones de marketing digital evolucionan con gran rapidez y aprovechar el momento clave es esencial. Por eso, la información debería estar actualizada al momento en todas las fuentes y mostrarse en el dashboard en tiempo real. (Ortiz, et, al; 2020).

3.5 Business Inteligencie y Big Data como estrategia del negocio.

Business Inteligencie

Es un término paraguas que abarca los procesos, las herramientas, y las tecnologías para convertir datos en información, información en conocimiento y planes para conducir de forma eficaz las actividades de los negocios. BI abarca las tecnologías de datawarehousing los procesos en el 'back end', consultas, informes, análisis y las herramientas para mostrar información (estas son las herramientas de BI) y los procesos en el 'front end'.

Componentes básicos de Business Inteligencie:

La principal razón de un proyecto de Business Intelligencie es el análisis de un problema o problemas interrelacionados.

• Problemática empresarial a la que queríamos dar respuesta.

• Un equipo de personas o una persona que lleve a cabo el análisis.

• Información de nuestros sistemas de pedidos y expediciones.

• Información externa de las tarifas de la empresa de transporte.

• Una base de datos a la que hemos llamado datawarehouse.

• Una aplicación de Business Intelligencie que nos permita trabajar con la información, analizarla y visualizar los resultados.

(Leopoldo, 2012)

Para (Arthur, 2013) el termino Big-Data se puede entender como la recolección de datos tanto de fuentes tradicionales como de fuentes digitales (no tradicionales) que representan una fuente para posteriores descubrimientos y análisis.

Analytics es un término nuevo que comienza a tomar sentido para las empresas, entendiéndose la misma, como el descubrimiento y la comunicación de patrones significativos de la información (data) o concebido como un método de análisis lógico de la información (Meier & Donze, 2012)

Entonces, con la anterior definición expuesta, ¿dónde radica la diferencia entre Business Analytics y Big-Data?, radica en las tres V's expuestas a continuación (McAfee & Brynjolfsson, 2012) y (Arthur, 2013):

- Volumen. - Es la cantidad de información recolectada, que incluye información de fuentes tradicionales y no tradicionales. Las empresas trabajan ahora con petabytes y exabytes.
- Velocidad. - Es la velocidad en la que la información es generada y fluye hacia la empresa. La velocidad en que la información es creada.
- Variedad. - Se refiere al tipo de información disponible para la empresa y para sus equipos de marketing.

Con las diferencias expuestas anteriormente se pueden resumir las ventajas y novedades del Big-Data en siete puntos (Capgemini, 2012):

1. Los volúmenes de información son mucho más grandes de lo que cualquier organización/empresa está acostumbrada a procesar.
2. Los volúmenes de información son mucho más amplios de lo que cualquier base de datos tradicional de una organización/empresa está acostumbrada a manejar.
3. La información externa es “traída” a la organización/empresa de terceras personas y fuentes públicas.
4. Alguna de la información proviene de las redes sociales.
5. Una cantidad significativa de la información puede ser altamente desestructurada (ej. Voz o video)
6. Varios conjuntos de información distintos están integrados conjuntamente para su análisis.
7. Análisis en tiempo real o cercano a tiempo real es requerido.

Referencias:

Arthur L. (2013). Big Data Marketing: Engage your customers more effectively and drive value. John Wiley & Sons, Inc. New Jersey: Estados Unidos.

Capgemini. (2012). Big Data: Next Generation Analysis.

Leopoldo M.C. (2012). BUSINESS INTELLIGENCE (BI).

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Business Review.

Meier, A., & Donze, L. (2012). Fuzzy Methods for Customer Relationship Management and Marketing. Business Science Reference.

3.6 Procesamiento de datos.

¿Qué es el procesamiento de datos?

Los datos pueden ser cualquier número o carácter que puede representar los valores de las mediciones o los fenómenos observables. Un solo dato es una medición de un fenómeno observable. La información medida es deducida algorítmicamente derivada y/o lógica y/o calculada estadísticamente a partir de múltiples datos. La información es una respuesta a una pregunta o un significativo estímulo que puede dar lugar a nuevas preguntas.

Por ejemplo, la recolección de datos sísmicos conduce a la alteración de los datos sísmicos para suprimir el ruido, mejorar la señal de los fenómenos sísmicos y migrar a la ubicación adecuada en el espacio. Suelen incluir los pasos de procesamiento de análisis de las velocidades y frecuencias, correcciones estáticas, entre otros. El procesamiento de datos sísmicos facilita una mejor interpretación, debido a que las estructuras del subsuelo y geometrías de reflexión son más evidentes.

Un bit es la unidad básica de almacenamiento de datos informáticos. Los bits se agrupan en bloques de ocho para formar bytes, que pueden almacenar un valor que la computadora puede interpretar como caracteres o calidad de la información. Un campo es un elemento de información que puede almacenarse. Son colecciones de bytes que almacenan las características de una entidad (para describir una entidad), por ejemplo, una persona, un ordenador, un coche etc. Un registro consta de dos o más valores o variables almacenados en posiciones consecutivas de memoria.

Un archivo está compuesto por una colección de registros. Una base de datos es una colección integrada de registros relacionados lógicamente, o archivos que consolidan los registros previamente almacenados en archivos separados en un fondo común de los registros de datos que proporciona datos para muchas aplicaciones. Procesamiento de datos se diferencia de la conversión de datos, cuando el proceso es simplemente para convertir datos a otro formato, y no implica ninguna manipulación o procesamiento de datos.

Diferencia entre Datos e Información

Los datos consisten en nada más que hechos (organizados o no organizados) que luego pueden ser manipulados en otras formas para que sean útiles y comprensibles, convirtiendo los datos en información.

El proceso de manipulación de hechos a información se conoce como "procesamiento". Para ser procesado por una computadora, los datos deben primero ser convertidos en un formato legible por máquina. Una vez que los datos están en formato digital, se pueden aplicar varios procedimientos sobre los datos para obtener información útil.

El procesamiento de datos puede involucrar varios procesos, incluyendo:

- Resumen de datos
- Agregación de datos
- Validación de datos
- Tabulación de datos
- Análisis estadístico

El procesamiento de datos puede o no puede distinguirse de la conversión de datos, que implica el cambio de datos en otro formato, y no implica ninguna manipulación de datos.

Durante el procesamiento, los datos brutos se utilizan como una entrada para producir información como una salida, normalmente en forma de informes y otras herramientas analíticas.

Etapas de procesamiento de datos

Recopilación de datos

La recopilación de datos es el primer paso en el procesamiento de datos. Los datos se obtienen de las fuentes disponibles, incluidos los archivos de testo y los almacenes de datos.

Es importante que las fuentes de datos disponibles sean confiables y estén bien construidas, por lo que los datos recopilados (y luego se utilizan como información) sean de la mejor calidad posible.

Preparación de datos

Una vez que se recopilan los datos, estos ingresan a la etapa de preparación de los datos. La preparación de datos, a menudo denominada "preprocesamiento", es la etapa en la que los datos sin procesar se limpian y organizan para la siguiente etapa de procesamiento de datos.

Durante la preparación, los datos sin procesar se verifican diligentemente para detectar cualquier error. El propósito de este paso es eliminar los datos incorrectos (datos redundantes, incompletos o incorrectos) y comenzar a crear datos de alta calidad para una mejor inteligencia empresarial.

Entrada de datos

Luego, los datos limpios se ingresan en su destino (tal vez un CRM como Salesforce o un almacén de datos), y se traducen a un idioma que se pueda comprender.

La entrada de datos es la primera etapa en la que los datos sin procesar comienzan a tomar la forma de información utilizable.

Procesamiento

Durante esta etapa, los datos ingresados en la computadora en la etapa anterior se procesan para su interpretación. El procesamiento se puede realizar mediante técnicas de filtrado, análisis y visualización de datos e incluso con algoritmos de aprendizaje automático, aunque el proceso en sí puede variar ligeramente dependiendo de la fuente de datos que se procesa (base de datos, redes sociales, dispositivos conectados, etc.) y su uso previsto (examen de patrones de publicidad, diagnóstico de dispositivos conectados, determinar las necesidades del cliente, etc.).

Interpretación de los datos

La etapa de salida es la etapa en la que los datos son finalmente utilizables para los usuarios. Se traducen los datos, se pueden leer, y a menudo en forma de gráficos, videos, imágenes, texto sin formato, etc.).

Los miembros de la empresa o institución ahora pueden comenzar a administrar los datos para sus propios proyectos de análisis de datos.

Almacenamiento de datos

La etapa final del procesamiento de datos es el almacenamiento. Una vez que se procesan todos los datos, se almacenan para su uso futuro. Si bien es posible que alguna información se use de inmediato, gran parte de ella tendrá un propósito más adelante.

Además, los datos almacenados correctamente son una necesidad para cumplir con la legislación de protección de datos como GDPR. Cuando los datos se almacenan correctamente, los miembros de la organización pueden acceder a ellos rápida y fácilmente cuando sea necesario.

Análisis de datos

Cuando el dominio desde el que se recogen los datos es una ciencia o ingeniería, el procesamiento de datos y de sistemas informativos se consideran términos demasiado amplios, y el término más especializado, análisis de datos se suele utilizar, centrándose en la altamente especializada y altamente precisa derivación algorítmica y cálculos estadísticos que se observan con menos frecuencia en el típico entorno empresarial. Las Mediciones de procesamiento de datos normalmente están representados por números enteros o de punto fijo con código binario o representaciones de los números decimales, mientras que la mayoría de las mediciones del análisis de datos son a menudo representados por representación de punto flotante de los números racionales.

Una vez que los datos están en formato digital, diversos procedimientos pueden aplicarse a los datos para obtener información útil. Procesamiento de datos pueden implicar diferentes procesos, entre ellos:

Entrada de datos

- Captura de datos
- Tipos de datos
- La depuración de los datos
- Integridad de los datos
- Codificación (cifrado) de datos
- Transformación de datos
- Traducción de datos
- Resúmenes de datos
- Agregación de datos
- Validación de datos
- Modelado de datos
- El análisis de datos
- El análisis de datos estadístico
- Visualización de datos
- Almacenamiento de datos
- Minería de datos
- Interpretación de datos

TIPOS DE DATOS

DATOS ESTRUCTURADOS

La mayoría de las fuentes de datos tradicionales son datos estructurados, datos con formato o esquema fijo que poseen campos fijos. En estas fuentes, los datos vienen en un formato bien definido que se especifica en detalle, y que conforma las bases de datos relacionales. Son los datos de las bases de datos relacionales, las hojas de cálculo y los archivos, fundamentalmente. Los datos estructurados se componen de piezas de información que se conocen de antemano, vienen en un formato especificado, y se producen en un orden especificado

DATOS SEMIESTRUCTURADOS

Los datos semiestructurados tienen un flujo lógico y un formato que puede ser definido, pero no es fácil su comprensión por el usuario. Datos que no tienen formatos fijos, pero contienen etiquetas y otros marcadores que permiten separar los elementos dato. La lectura de datos semiestructurados requiere el uso de reglas complejas que determinan cómo proceder después de la lectura de cada pieza de información.

DATOS NO ESTRUCTURADOS

Los datos no estructurados son datos sin tipos predefinidos. Se almacenan como “documentos” u “objetos” sin estructura uniforme, y se tiene poco o ningún control sobre ellos. Datos de texto, video, audio, fotografía son datos no estructurados.

PROFESIONALES DE ANÁLISIS DE DATOS: ANALISTAS Y CIENTÍFICOS DE DATOS

Existe una enorme escasez de especialistas en gestión y análisis de datos. IDC, Gartner, Forrester, McKinsey, las grandes consultoras tecnológicas reconocen en sus últimos estudios sobre Big Data, que las empresas y organizaciones no disponen de suficiente talento para afrontar los retos tecnológicos y organizativos. Solo en los Estados Unidos, estadísticas fiables confirman que se necesitarán entre 140.000 y 190.000 expertos en datos hasta 2018: estadísticos, matemáticos, analistas, directivos (managers) con una experiencia híbrida en negocios y proyectos cuantitativos, y expertos en software y en lenguajes de programación específicos de análisis de datos. Además, el informe de McKinsey (2011), señala la necesidad de disponer de 1,5 millones de gerentes alfabetizados en análisis de datos.

3.7 Arquitectura Big Data

El termino Big Data cada día que pasa, se torna más importante, es por esto que en la presente investigación se estudia, analiza y da a conocer de manera exhaustiva las diferentes arquitecturas de Big Data, con sus características, herramientas, tecnologías y estándares relacionadas con dicho termino, con el único objetivo de brindar una ayuda al sector empresarial para una posible implementación de Big Data.

Big Data refiere a la información que no puede ser procesada o analizada mediante procesos tradicionales.

Para zdnet.com (2010), Big Data son “cantidades masivas de datos que se acumulan con el tiempo que son difíciles de analizar y manejar utilizando herramientas comunes de gestión de bases de datos”.

Según Dans (2001) Big Data también alude “al tratamiento y análisis de enormes repositorios de datos, tan desproporcionadamente grandes que resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales”

Para el analista Dan Kusnetzky, del Grupo Kusnetzky (Preimesberger, 2011) “las herramientas, procesos y procedimientos que permitan a una organización crear, manipular y administrar grandes conjuntos de datos e instalaciones de almacenamiento”

Por último, para Krishnan (2013), Big Data son grandes volúmenes de datos disponibles con desiguales grados de complejidad, con diferentes velocidades y con

gran ambigüedad. Datos que no se pueden procesar con tecnologías tradicionales, además, son generados desde redes sociales, redes de sensores, dispositivos de rayos X, centrales nucleares, aviones, ventas, marketing, compras y finanzas personales.

Es de aclarar que los metadatos son información que describe características de cualquier dato, como el nombre, la ubicación, la importancia percibida, la calidad y sus relaciones con otros objetos de datos que la empresa considere digno de la gestión (Soares, 2012b)

Arquitectura De Big Data

Existen varios modelos de arquitectura de Big Data, de los cuales se mostrarán tres:

Arquitectura de procesamiento de Big Data propuesta por Krishnan

La arquitectura presentada por Krishnan (2010) es similar al tratamiento de la gestión de datos que hace bastante tiempo se conoce, el cual consiste en cuatro etapas: recolección o recopilación, carga, transformación y extracción de datos. Para este caso, el autor lo llama enfoque de procesamiento de Big Data (Demchenko, Ngo y Membrey, 2012

Etapa 1. Recolección o recopilación o fuentes de datos de Big Data: 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

Redes sociales. De medios o redes como Facebook, Twitter, LinkedIn y blogs. Por ejemplo, como menciona Soares

Generados por el ser humano. Son datos producto de grabaciones de voz, correo electrónico, documentos en papel, las encuestas y registros electrónicos

También se consideran fuente de datos: “registros, trayectorias de navegación, datos de redes sociales, transferencias de noticias, emails, salida

Transacciones. Son datos que pueden provenir de registros detallados de llamadas (charging data record, CDR) de las telecomunicaciones, registros de facturación de servicios públicos que están cada vez más disponibles en formatos semiestructurados y no estructurados.

Etapa 2. Carga de datos de Big Data

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 (Krishnan, 2013)

Es de aclarar que los metadatos son información que describe características de cualquier dato, como el nombre, la ubicación, la importancia percibida, la calidad y

sus relaciones con otros objetos de datos que la empresa considere digno de la gestión

Se busca vincular los datos entre el conjunto de datos estructurados y no estructurados con metadatos y datos maestros. Se deben transformar los datos no estructurados en estructurados. Es importante acudir a la integridad referencial, la cual ayuda inicialmente con la clave principal y las demás relaciones en una base de datos tradicional

Etapa 3. Transformación de datos de Big Data

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.

Etapa 4. Extracción de datos de Big Data

El objetivo de la extracción es obtener datos para su posterior análisis, generar informes operativos y su posible visualización y, por último, y no el más importante, para su almacenamiento (Krishnan, 2013).

Arquitectura de Big Data propuesta por Bob Marcus

G. Servicios de Apoyo

F. Aplicaciones e interfaces de usuario

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 externo

Fuentes de datos externas

El componente A, es parte de la arquitectura de datos que suministra las entradas de datos externas y la producción de los componentes internos de Big Data

Secuencia y procesamiento ETL

Las tareas que se desarrolla en el componente B son filtrar y transformar los flujos de datos provenientes de los recursos externos. El procesamiento de datos que se realiza es el llamado “en movimiento” entre los almacenes de datos.

Fundación altamente escalable

• El primero, a nivel de la infraestructura, existe con el fin de poder atender el almacenamiento y procesamiento de grandes volúmenes de datos.

• El segundo se refiere 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 [usando componentes menos caros puede apoyar el crecimiento ilimitado de almacenamiento de datos”.

Bases de datos operacionales y de Analíticas

En el componente D se proponen dos 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 desnormalización).

• Bases de datos operacionales. Estas bases de datos mantienen una excelente operación en lectura y escritura en general de forma eficiente.

Analítica e interfaces de bases de datos

• 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

• Análisis de interfaces interactivas. “Los almacenes de datos pueden ser bases de datos escalables horizontalmente sintonizados para las respuestas interactivas

Aplicaciones e interfaz de usuario

El componente F se refiere a las aplicaciones e interfaces de usuario, las cuales no deben ser algoritmos complejos, al usar grandes cantidades de datos distribuidos.

Servicios de apoyo

Es el componente G. Estos servicios hacen referencia a los componentes necesarios para la implementación y gestión de sistemas robustos de Big Data, los cuales se pueden discriminar así:

• Diseñar, desarrollar e implementar herramientas. Consiste en tener a la mano herramientas bien desarrolladas, es decir, de alto nivel de calidad, de manera que sirvan para implementar soluciones Big Data.

• Seguridad. Aspecto importante a nivel de controles de seguridad de grandes volúmenes de datos, pues en la actualidad son escasos o limitados. Se busca ampliar la seguridad en el Big Data.

• Gestión de procesos. Los distribuidores comerciales son el suministro de herramientas de gestión de procesos para aumentar las implementaciones de código abierto

• Gestión de recursos de datos. En se hace hincapié en Herramientas de control de datos de código abierto que son todavía inmaduras. Estos serán aumentados en un futuro por los proveedores comerciales.

3.8 Learning BigData

Para empezar, se muestran definiciones sobre el concepto de Learning BigData:

Según (Lazcano, s/f). Grandes datos. Este sistema se encarga de procesar y analizar grandes cantidades de datos. Pero no nos referimos a ningún número. Estos números pueden estar en el rango de millones de gigabytes de información, claramente cantidades que no pueden ser absorbidas por los métodos de procesamiento tradicionales.

Además de la cantidad de almacenamiento de datos, la utilidad de Big Data se puede resumir en sus otras 2 V: velocidad y variedad. En segundos, se están creando más y más datos a velocidades inmanejables; por otro lado, existe heterogeneidad en las características de esta información: tamaño, tipo, formato, estructura y múltiples fuentes.

Según (Lazcano, s/f). La relación del big data con machine learning y deep learning se produce a raíz de sus mismas diferencias. Es decir, aquello que los diferencia es el mismo motivo por el cual se complementan como técnicas informáticas.

En este sentido, tenemos que:

El big data extrae y procesa los datos para disponibilizarlos ante los algoritmos de machine learning. Se puede decir que el big data es la fuente de ingesta de datos para el ML y DL.

El machine learning toma los datos procesados por el big data y los analiza para generar insights de negocio o aprender a realizar ciertas tareas automáticamente.

El deep learning ingiere los datos más importantes del big data para aprender sobre ellos a niveles mucho más profundos y para realizar tareas más complejas. (Lazcano, 2019).

El Big Data y la gran cantidad de información contenida en él ha permitido que las máquinas adquieran una mayor importancia, lo que tiene como resultado la generación de herramientas de Inteligencia Artificial y Machine Learning.

El Big Data se está moviendo rápidamente hacia una nueva etapa de madurez que promete un impacto aún mayor en los negocios, así como también una disrupción en la industria en los próximos años. A medida que las iniciativas maduran, las organizaciones combinan ahora la agilidad de los procesos de Big Data con la escalabilidad de la inteligencia artificial. Estas dos características ayudan a acelerar la entrega de valor aún más que antes.

Por otro lado, la habilidad de gestionar grandes volúmenes y fuentes de datos está habilitando las capacidades de la IA y, especialmente de machine learning, que permanecían dormidas desde hace décadas debido a tres problemas relativamente frecuentes:

La falta de disponibilidad de datos que sufrían las empresas.

Los tamaños de muestra limitados, que afectaban negativamente a las capacidades de las organizaciones.

La imposibilidad de analizar cantidades masivas de datos en milisegundos, por falta de herramientas adecuadas.

Esta última abre la posibilidad a que las máquinas “aprendan” por medio de la generación de algoritmos, sin que alguien tenga que programarlas. El Machine Learning contempla cuatro enfoques principales:

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje semi supervisado
- Aprendizaje de reforzamiento

Entre los principales beneficios del Machine Learning se encuentra la identificación de tendencias en el comportamiento de los clientes, patrones operativos en las empresas y bases sólidas para la generación de productos, anticipándose a las necesidades futuras del mercado. (Bello, 2021).

Las grandes compañías se encuentran aún en la tarea de absorber tanto conocimiento como sea posible a partir del Big Data y del Machine Learning para obtener el mayor provecho de ambas, siendo uno de los principales retos la selección y la filtración de los datos para crear un panorama de posibilidades más eficiente y eficaz. (Bello, 2021).

Data Mining

Según (Bello, 2021). El minado de datos es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos.A pesar de que la idea del Data Mining puede parecer una innovación tecnológica muy reciente, en realidad este término apareció en los años sesenta conjuntamente con otros conceptos como, por ejemplo, el data fishing o data archeology. No obstante, no fue hasta los años ochenta cuando empezó su consolidación.

La minería de datos surgió con la intención o el objetivo de ayudar a comprender una enorme cantidad de datos y que estos pudieran ser utilizados para extraer conclusiones para contribuir en la mejora y el crecimiento de las empresas. Sobre todo, por lo que hace a las ventas o fidelización de clientes.Su principal finalidad es explorar, mediante la utilización de distintas técnicas y tecnologías, bases de datos enormes de manera automática. El objetivo es encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos que se han ido recopilando con el tiempo. Estos patrones pueden encontrarse utilizando estadísticas o algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales. (Bello, 2021).

Por tanto, los datos son el medio o la base para llegar a conclusiones y transformar estos datos en información relevante, para que las empresas puedan abarcar mejoras y soluciones que les ayuden a conseguir sus objetivos.

Social Mining

Según (Redacción, 2016). El Data Mining o minería de datos centra su estudio en identificar patrones de comportamiento, a partir del análisis de grandes volúmenes de datos. Entre los sectores donde más se aplica el Data Mining se encuentra la banca, pero no es el único. La minería de datos puede servir para detectar los movimientos sociales en la red, aunque a esta estrategia se la conoce como Social Media Mining.

El Social Media Mining o minería de los medios de comunicación, es un proceso que trata de analizar y extraer patrones de conducta a partir de los datos que proporcionan las redes sociales.Desde los primeros días de la humanidad, las normas sociales se han establecido en los grupos sociales. Desde la antigua Roma hasta la era digital el comportamiento humano ha sido caso de estudio. En la actualidad, las redes sociales son la muestra y escaparate para observar, analizar e identificar los patrones de comportamiento. Aquí nace el Social Media Mining, que pretende analizar y extraer patrones significativos de comportamiento, a partir de la interacción, comentarios y preferencias de los usuarios de redes sociales.

Además de la informática y diversos algoritmos que ayudan a recopilar toda la información, se aplican otras disciplinas. La sociología, la etnografía, la estadística, optimización y las matemáticas permiten que se deduzcan patrones de comportamiento dentro de las redes sociales. (CulturaCRM., 2016

Zdnet.com. (2010). [En línea]. Recuperado de: http://www.zdnet.com/search?q=big+data [Consultado el 29 de abril de 2022]

Preguntas referentes a la Unidad

1. El diseño del estudio tiene 4 características clave:
    a. Neutralidad, Confiabilidad, Validez, Generalización
    b. Neutralidad, Disponibilidad, Validez, Generalización
    c. Neutralidad, Confiabilidad, Variedad, Generalización
2. Se define como la elección de métodos y técnicas por parte del investigador para combinarlos de manera lógica:
    a. Diseño de la Investigación
    b. Metodología de la Investigación
    c. Investigación
3. Suele expresar mediante “Y”, es el efecto que se produce a partir de la variable independiente.
    a. Variables dependientes
    b. Variables
    c. Variables independientes
4. Tipos de diseños de investigación.
    a. Cuantitativo y Cualitativo
    b. Completo y Cualitativo
    c. Cuantitativo y Confiable
5. El procesamiento de datos puede involucrar varios procesos, excepto:
    a. Manipulación de datos
    b. Análisis estadístico
    c. Tabulación de datos
    d. Resumen de datos
6. La minería de datos puede servir para detectar los movimientos sociales en la red:
    a. True
    b. False
7. La sociología, la etnografía, la estadística, optimización y las matemáticas permiten que se deduzcan patrones de comportamiento dentro de las redes sociales.
    a. True
    b. False
8. El minado de datos es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática.
    a. True
    b. False
9. La relación del big data con machine learning y deep learning se produce a raíz de sus mismas diferencias.
    a. True
    b. False
10. El procesamiento de los datos diferencia de la conversión de datos, cuando el proceso es simplemente para convertir datos a otro formato, y no implica ninguna manipulación o procesamiento de datos.
    a. True
    b. False

Respuestas correctas: a

REFERENCIAS:

Ali, H. S., Arshad, M. J., & Sumra, I. A. (2019). 7, Vs of Big Data: A Survey.

Ángel M. Rayo. (2016). Tipos de datos en Big Data: clasificación por categoría y por origen. Sitio web: https://netmind.net/es/tipos-de-datos-en-big-data-2/

Argonza, J. S. (2012). Dispositivos para el almacenamiento de grandes volúmenes de información "Big data". Obtenido de Google schoolar en: Obtenido de Google schoolar en: https://bit.ly/38Bnjxl

Ariel Ortiz Ramírez. (2010). Python como primer lenguaje de programación. 19-05-2010, de Tecnológico de Monterrey, Campus Estado de México Sitio web: https://arielortiz.info/publicaciones/primer_lenguaje_30_jun_2010.pdf

Calles, A. (2022, 3 enero). Amazon Transporte: ¿Por qué su gestión es un gran ejemplo? Drivin. https://blog.driv.in/es/amazon-transporte-por-que-su-gestion-es-un-gran-ejemplo/

Caminero Herráez, A.C & Grau Fernández, L. (2016). *INTRODUCCIÓN AL MANEJO DE DATOS MASIVOS CON HADOOP.* UNED. Recuperado de: https://www.cartagena99.com/recursos/alumnos/apuntes/Practica-SBD-2015-16_v1.pdf

del Conocimiento, I. D. I. (2020, 28 agosto). Las 7 V del Big data: Características más importantes. Instituto de Ingeniería del Conocimiento. https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/

Equipo de edX. (2021). R vs. Python para la ciencia de datos: Explicación y consejos de aprendizaje. 19-05-2022, de Desconocido Sitio web: https://blog.edx.org/es/r-vs-python-para-la-ciencia-de-datos-explicacion-y-consejos-de-aprendizaje#:~:text=Python%20es%20mucho%20m%C3%A1s%20sencillo,m%C3%A1s%20sencilla%20para%20el%20an%C3%A1lisis.

García, A. (2012). *manual práctico de SQL.* {lwp} Comunidad de programadores. Recuperado de: https://www.lawebdelprogramador.com/cursos/archivos/ManualPracticoSQL.pdf

Introducción a Hadoop y su ecosistema | Dataprix TI. (2013, 3 abril). Dataprix. https://www.dataprix.com/es/blog-it/jcasanella/introduccion-hadoop-y-su-ecosistema

Joyanes L. (2013). Big Data: Análisis de grandes volúmenes de datos en organizaciones. México: Alfaomega.

Méndez A. (2012) "Big Data: ¿humo o reto corporativo?" www.penteo.com [Consultado: 18 de Mayo de 2022]. Disponible en Internet: https://bit.ly/3yKEU0r

Pérez, M. (2015). *BIG DATA-Técnicas, herramientas y aplicaciones*. Alfaomega Grupo Editor. Obtenido de Google schoolar en: https://bit.ly/3PjkMIV

UNIDAD IV. Aplicaciones, plataformas y tecnologías en Big data.
OBJETIVO: Aplicar plataformas, software en el procesamiento de información en BIG DATA.

4.1. Sistemas operativos, servidores y redes.

El análisis de Big Data es una referencia a sistemas que manipulan grandes conjuntos de información, presentando dificultades habituales dentro de la captura, búsqueda, compartición análisis y visualización de los datos. Dado al gran avance que existe hoy en día en las tecnologías de la información, se han tenido que enfrentar nuevos desafíos. Y los avances en hardware y software han ayudado a disminuir este tipo de problemas. ***(Pérez, 2015)***

Siguiendo a ***Argonza J. en "Dispositivos para el almacenamiento de grandes volúmenes de información Big data" (2012***), en la actualidad se identifican cuatro tecnologías diferentes para el almacenamiento de grandes volúmenes información en las siguientes;

a. Dispositivos SAN (Storage Area Network, Red de Área de Almacenamiento), es una red de alta velocidad, dedicada especialmente para el almacenamiento de datos y que está conectada a uno o más servidores a través de fibra óptica. Las redes SAN comúnmente utilizan arreglos RAID (Redundant Array of Independent Disks), "Conjunto Redundante de Discos Independientes" de tecnología SCSI, aunque también pueden utilizan otros dispositivos de almacenamiento como discos ópticos y cintas.
b. Dispositivos NAS (Network Attached Storage o Almacenamiento Conectado a Red), es un dispositivo capaz de compartir su capacidad de almacenamiento a través de una red (normalmente vía TCP/IP), haciendo uso de un sistema operativo optimizado que emplea los protocolos CIFS, NFS, FTP o TFTP.

   Los protocolos de comunicaciones NAS, están basados en archivos, por lo que un cliente solicitará el archivo completo al dispositivo de almacenamiento y lo manejará localmente.

c. Dispositivos SAN/NAS ofrecen a los usuarios una solución híbrida SAN-NAS que permite dar una respuesta integral a sus necesidades de almacenamiento en una única fuente de almacenamiento. Este tipo de solución permite distribuir parte de la capacidad de almacenamiento en solo una unidad para ser utilizado por aplicaciones de archivo NAS, mientras otras partes de la misma unidad de almacenamiento pueden ser utilizadas como SAN.
d. Dispositivos de almacenamiento orientado a objetos, los dispositivos de almacenamiento por objetos constituyen un nuevo tipo de periféricos diseñados para ofrecer acceso, almacenamiento, protección y distribución inteligentes de contenido digital fijo. Este tipo de equipos ofrecen la posibilidad de almacenar desde unos pocos terabytes a varios petabits de información.

   Este tipo de equipos basados en objetos ofrece una solución para el rápido crecimiento de datos no estructurados como archivos, imágenes y videos que llegan a congestionar el almacenamiento principal, complicando la administración de los datos.

En cuanto a la parte de los sistemas operativos, **Pérez (2015)** menciona que realmente hoy en día existe una gran compatibilidad con los sistemas actuales, independientemente si sean distribuciones de Linux, Windows o Mac, incluso actualmente se han desarrollado aplicaciones para móviles (android y apple).

Según Joyanes L (2013) hace mención que cloud computing ofrecen tres modelos de servicio que se ofertan a los clientes y usuarios de la nube para organizaciones, empresas y usuarios son:

Software como servicio. Al usuario se le ofrece la capacidad de que las aplicaciones que su proveedor le suministra corran en una infraestructura de la nube, siendo dichas aplicaciones accesibles a través de una interfaz del cliente tal como un navegador Web (correo electrónico Web, Gmail o Yahoo) o una interfaz de programa. El usuario carece de cualquier control sobre la infraestructura de la nube, como servidores, sistemas operativos, almacenamiento, incluso sobre las propias

aplicaciones, excepto por las posibles configuraciones de usuario o personalizaciones que se le permitan realizar.

· Plataforma como servicio. Al usuario se le permite desplegar aplicaciones propias (ya sean adquiridas o desarrolladas por el propio usuario) creadas utilizando lenguajes y herramientas de programación soportadas por el proveedor. El consumidor no administra ni controla la infraestructura de la nube, incluyendo redes, servidores, sistemas operativos ni almacenamiento de su proveedor, que es quien ofrece la plataforma de desarrollo y las herramientas de programación. El usuario tiene control sobre las aplicaciones desplegadas, y es quien mantiene su control, aunque no de toda la infraestructura subyacente.

· Infraestructura como servicio. El proveedor ofrece al usuario recursos como capacidad de procesamiento, de almacenamiento, comunicaciones y otros recursos de computación donde el consumidor es capaz de desplegar y ejecutar software específico que puede incluir sistemas operativos y aplicaciones. El consumidor no administra ni controla la infraestructura fundamental de la nube, pero tiene control sobre sistemas operativos, almacenamiento, aplicaciones desplegadas; y, posiblemente, un control limitado de componentes seleccionados de redes (cortafuegos de los hospedajes, host firewalls).

Algunos ejemplos de plataformas mostradas por ***Joyanes (2013)***

DynamoDB Amazon Web Services (AWS) anunció, a primeros de 2012, el lanzamiento de DynamoDB, una base de datos NoSQL que Amazon ha desarrollado y probado internamente durante los años anteriores. Amazon ha lanzado esta solución al comprobar el auge creciente del mercado de los grandes volúmenes de datos y el número de aplicaciones Web. Esta oferta se une a su base de datos relacional Amazon RDS, y a otra base de datos NoSQL denominada SimpleDB, ambas con gran experiencia en el mercado desde sus herramientas de cloud computing.

ParAccel Analytic Database (PADB) que es una plataforma de bases de datos analítica de procesamiento paralelo masivo (MPP) y columnar (por columnas) con

características muy potentes para optimización y compilación de consultas, compresión e interconexión de redes. PADB es desplegable en entornos empresariales incluso en otros entornos operativos estándares como cloud computing y virtualización.

4.2 Lenguajes de programación Python y R.

Python es un lenguaje de programación moderno creado por Guido van Rossum a inicios de los años noventa. La implementación canónica, conocida como CPython, está bajo una licencia de software libre y se puede descargar del sitio oficial. El que sea una tecnología abierta y libre tiene ventajas importantes sobre tecnologías propietarias. La principal es que se puede usar sin tener que cubrir costos de licencias. Esto quiere decir que un alumno puede seguir usando Python gratuitamente fuera de una institución, por ejemplo, para escribir software en un entorno comercial, o para continuar con sus estudios de posgrado en alguna otra universidad. Ariel Ortiz Ramírez. (2010).

Características de Python

1. Es orientado a objetos, se soporta también los estilos de programación procedural y funcional.
2. Corre en múltiples plataformas, incluyendo Windows, Mac OS y Linux.
3. Su sintaxis y semántica es sencilla y consistente.
4. Utiliza tipos dinámicos.
5. Es adecuado tanto para programar scripts como aplicaciones de gran tamaño.
6. Es muy modular
7. Cuenta con administración automática de memoria a través de recolección de basura.
8. Incluye una poderosa y extensa biblioteca de clases. Ariel Ortiz Ramírez. (2010).

Python y R se consideran lenguajes de programación esenciales para la ciencia de datos. Lo ideal sería dominar ambos para tener una base de programación completa.

¿Cuál es la diferencia entre Python y R?

Mientras que el lenguaje R es más especializado, Python es un lenguaje de programación de propósito general diseñado para una variedad de casos de uso.

Hay similitudes entre los dos lenguajes, por lo que la experiencia en uno de ello puede ser útil para el otro, por ejemplo: tanto Python como R son populares lenguajes de programación de código abierto respaldados por prósperas

comunidades. Ambos pueden practicarse también en el entorno agnóstico del lenguaje.

Python: El lenguaje de programación para todos los propósitos

Según los datos de Stack Overflow, Python es el lenguaje de programación de más rápido crecimiento en todo el mundo. Es muy accesible para los principiantes y ofrece el tipo de versatilidad que los desarrolladores web necesitan para crear sitios web tan variados como Spotify, Instagram, Reddit, Dropbox y el Washington Post. ¿No sabes cómo usar un caret o qué es una regresión? Python será el punto de partida más amigable para ti.

Python es un lenguaje de programación orientado a objetos, como Javascript o C++, que proporciona estabilidad y modularidad a los proyectos, sin importar su tamaño. Ofrece un enfoque flexible para el desarrollo web y la ciencia de datos que se siente intuitivo incluso si nunca has aprendido un lenguaje de programación antes.

"3 razones para aprender a programar en R para la ciencia de datos"

R no es un lenguaje de propósito general, pero dependiendo de dónde o cómo planees trabajar, podría ofrecer muchas ventajas que no están disponibles con un lenguaje de propósito general.

1. R está construido para la estadística: El análisis estadístico robusto es posible con Python, pero no tendrás las bibliotecas y funciones específicas de la sintaxis como lo haces con R. El lenguaje hace que sea mucho más intuitivo construir y comunicar los resultados de estos tipos específicos de programas. Los estadísticos y los analistas de datos utilizan R para manejar grandes conjuntos de datos con mayor facilidad utilizando modelos de aprendizaje automático estándar y minería de datos.

2. R es académico: R es casi la elección por defecto para trabajar en el mundo académico. R es muy adecuado para un subcampo del aprendizaje automático conocido como aprendizaje estadístico. Cualquier persona con una formación formal en estadística debería reconocer la sintaxis y la construcción de R.

3. R es intuitivo para el análisis: Puede que R no funcione con una gran variedad de proyectos, pero es la mejor opción para el trabajo de análisis e inferencia. Si piensa trabajar en un campo especializado, querrá un lenguaje de programación especializado. R también ofrece un potente entorno ideal para los tipos de visualizaciones de datos que emplean los científicos de datos.

Es mejor elegir Python si:	Es mejor elegir R si:
• No tiene experiencia en programación • El objetivo principal es la producción o el despliegue • Quieres construir nuevos modelos desde cero • El código de los proyectos debe ser legible.	• Tiene previsto trabajar en la investigación o en el mundo académico • El trabajo tiene un fuerte componente estadístico y de análisis • Deseas hacer uso de amplias bibliotecas para soluciones existentes • Deseas hacer uso de amplias bibliotecas para soluciones existentes • Las características específicas de la sintaxis son importantes • La comunicación de resultados complejos es clave

Tabla1. Tabla de elección de Python o R.

4.3 Almacenamiento y procesamiento de la información en Big Data.

Uno de los objetivos del uso de las tecnologías Big Data es el de transformar los datos en conocimiento útil para la empresa, y para ello se necesitan herramientas Big Data que nos ayuden a analizar, procesar y almacenar todos los datos recogidos. Un gran número de entre las mejores herramientas usadas en Big Data son open source, lo que da fe del éxito de este modelo de desarrollo, además de las alternativas de pago.

Big Data (Concepto).

Según Merv (2011). define este término como: “Big Data excede el alcance de los entornos de hardware de uso común y herramientas de software para capturar, gestionar y procesar los datos dentro de un tiempo transcurrido tolerable para su población de usuarios.

Otro termino redactado por McKinsey (2011). “Big Data se refiere a los conjuntos de datos cuyo tamaño está más allá de las capacidades de las herramientas típicas de software de bases de datos para capturar, almacenar, gestionar y analizar”.

4.3.1 Las 7Vs.

Características.

Las características más importantes del Big Data perfectamente se pueden clasificar en cuatro magnitudes, más conocidas como las cuatro V del Big Data, relativas a volumen, variedad, velocidad y veracidad. A estas cuatro V, podemos añadir tres más, como pueden ser la de Viabilidad y Visualización. Pero si hablamos de V en Big Data no podemos dejar pasar la principal característica del análisis de datos que es la V de Valor de los datos. del Conocimiento. (2020).

1. *Volumen de información.*

El volumen se refiere a la cantidad de datos que son generados cada segundo, minuto y días en nuestro entorno. Es la característica más asociada al Big Data, ya que hace referencia a las cantidades masivas de datos que se almacenan con la finalidad de procesar dicha información, transformando los datos en acciones.

2. *Velocidad de los datos.*

La velocidad se refiere a los datos en movimiento por las constantes interconexiones que realizamos, es decir, a la rapidez en la que son creados, almacenados y procesados en tiempo real.

Para los procesos en los que el tiempo resulta fundamental, tales como la detección de fraude en una transacción bancaria o la monitorización de un evento en redes sociales, estos tipos de datos deben estudiarse en tiempo real para que resulten

útiles para el negocio y se consigan conclusiones efectivas. del Conocimiento. (2020).

3. *Variedad de los datos.*

La variedad se refiere a las formas, tipos y fuentes en las que se registran los datos. Estos datos pueden ser datos estructurados y fáciles de gestionar como son las bases de datos, o datos no estructurados, entre los que se incluyen documentos de texto, correos electrónicos, datos de sensores, audios, vídeos o imágenes que tenemos en nuestro dispositivo móvil, hasta publicaciones en nuestros perfiles de redes sociales, artículos que leemos en blogs, las secuencias de click que hacemos en una misma página, formularios de registro e infinidad de acciones más que realizamos desde nuestro Smartphone, Tablet y ordenador.

4. *Veracidad de los datos.*

Cuando hablamos de veracidad nos referimos a la incertidumbre de los datos, es decir, al grado de fiabilidad de la información recibida. Es necesario invertir tiempo para conseguir datos de calidad, aplicando soluciones y métodos que puedan eliminar datos imprevisibles que puedan surgir como datos económicos, comportamientos de los consumidores que puedan influir en las *decisiones de compra.*

5. *Viabilidad.*

La inteligencia empresarial es un componente fundamental para la viabilidad de un proyecto y el éxito empresarial. Se trata de la capacidad que tienen las compañías en generar un uso eficaz del gran volumen de datos que manejan. La inteligencia competitiva también se asocia con la innovación de los equipos de trabajo y el uso de tecnologías empleadas. Una empresa inteligente analiza, selecciona y monitoriza la información con el fin de conocer mejor el mercado en el que opera, a sus clientes y diseñar estrategias eficaces.

6. *Visualización de los datos.*

Cuando hablamos de visualización nos referimos al modo en el que los datos son presentados. Una vez que los datos son procesados (los datos están en tablas y hojas de cálculo), necesitamos representarlos visualmente de manera que sean legibles y accesibles, para encontrar patrones y claves ocultas en el tema a investigar. Para que los datos sean comprendidos existen herramientas de visualización que te ayudarán a comprender los datos gráficamente y en perspectiva contextual.

7. *Valor de los datos.*

El dato no es valor. Tampoco tienes valor por el mero hecho de recopilar gran cantidad de información. El valor se obtiene de datos que se transforman en información; esta a su vez se convierte en conocimiento, y este en acción o en

decisión. El valor de los datos está en que sean accionables, es decir, que los responsables de las empresas puedan tomar una decisión (la mejor decisión) en base a estos datos.

4.3.2 Tipos de Datos para Almacenar (Multivariedad de datos en la BigData). Tipos de datos de Big Data

La categorización de los datos es importante para cualquier proyecto, y en especial cuando vamos a trabajar con grandes volúmenes (Big Data).

Dos de las categorizaciones más utilizadas en Big Data suelen ser las que relacionan la estructura de los datos y las que dependen del origen de los mismos:

Tipos de datos de Big Data por categorías

Los tipos de datos se suelen organizar en 2 categorías principales:

1. Datos Estructurados.

- Creados: datos generados por nuestros sistemas de una manera predefinida (registros en tablas, ficheros XML asociados a un esquema)
- Provocados: datos creados de manera indirecta a partir de una acción previa (valoraciones de restaurantes, películas, empresas (Yelp, TripAdvisor, …)
- Dirigido por transacciones: datos que resultan al finalizar una acción previa de manera correcta (facturas autogeneradas al realizar una compra, recibo de un cajero automático al realizar una retirada de efectivo, …)
- Compilados: resúmenes de datos de empresa, servicios públicos de interés grupal. Entre ellos nos encontramos con el censo electoral, vehículos matriculados, viviendas públicas, …)
- Experimentales: datos generados como parte de pruebas o simulaciones que permitirán validar si existe una oportunidad de negocio.

2. No estructurados:

- Capturados: datos creados a partir del comportamiento de un usuario (información biométrica de pulseras de movimiento, aplicaciones de seguimiento de actividades (carrera, ciclismo, natación, …), posición GPS)
- Generados por usuarios: datos que especifica un usuario (publicaciones en redes sociales, vídeos reproducidos en Youtube, búsquedas en Google, …)
- Multi-estructurados o híbridos:
- Datos de mercados emergentes
- E-commerce
- Datos meteorológicos

M. Rayo. (2016).

4.3.3Sistemas de almacenamiento distribuido.

El análisis y gestión de datos se ha convertido en una verdadera necesidad para las empresas del siglo XXI que quieran adaptarse a los cambios del contexto digital. Esto es debido al gran volumen de datos que se generan diariamente.

HDFS

HDFS responde a las siglas Hadoop Distributed File System, es decir, un sistema de archivos distribuido que usa Hadoop como sistema de almacenamiento de ficheros. Está formado por clústeres GNU/Linux y construido en Java, aspecto que permite instalarlo en multitud de máquinas.

La arquitectura de HDFS se basa en dos componentes fundamentales:

Namenode: este nodo actúa como servidor maestro y se encarga principalmente de regular el acceso de los clientes a los ficheros, mantener en memoria la metadata del sistema de ficheros y controlar los bloques que tiene cada datanode.

Datanode: son los encargados de almacenar y recuperar los bloques. Estos nodos formarán un cluster cuyos objetivos serán realizar operaciones de lectura y escritura de los archivos a raíz de las peticiones de los clientes. *Redacción España (2019).*

4.3.4 Introducción Hadoop. Arquitectura Hadoop.

Introducción a Hadoop.

Hadoop es una plataforma muy común para trabajar con datos, pero puede ser un poco difícil entender exactamente qué es y qué hace.

Hadoop era el nombre del peluche del hijo de uno de sus desarrolladores. Era un elefante; de ahí su logo. ¿Qué es Hadoop y qué hace? Lo más importante es que Hadoop no es una única cosa. Es una colección de aplicaciones de software para trabajar con big data. Es un FrameWork o una plataforma que consta de varios módulos. Quizás la parte más importante de Hadoop es el sistema de archivos distribuidos (o HDFS), que toma partes de información, una colección de información, y repartirla entre varias computadoras.

Hadoop puede tener cientos o millones de archivos separados repartidos entre esas computadoras, todos conectados entre ellos mediante el software.

Hadoop es de código abierto.

Hadoop es extremadamente popular en el mundo del big data, y hay un desarrollo muy activo de Hadoop, pero también una fuerte competencia en el mercado. No

todos están dispuestos a dejar que Hadoop monopolice el mercado. (Poulson, 2019).

Arquitectura.

¿Qué es Hadoop?

Apache Hadoop es un framework que permite el procesamiento de grandes volúmenes de datos a través de clusters, usando un modelo simple de programación. Además, su diseño permite pasar de pocos nodos a miles de nodos de forma ágil. Hadoop es un sistema distribuido usando una arquitectura Master-Slave, usando para almacenar su Hadoop Distributed File System (HDFS) y algoritmos deMapReduce para hacer cálculos (*Dataprix*, 2013).

El sistema de ficheros HDFS

HDFS es el sistema de almacenamiento, es un sistema de ficheros distribuido. Fue creado a partir del Google File System (GFS). HDFS se encuentra optimizado para grandes flujos y trabajar con ficheros grandes en sus lecturas y escrituras. Su diseño reduce la E/S en la red. La escalabilidad y disponibilidad son otras de sus claves, gracias a la replicación de los datos y tolerancia a los fallos. (*Dataprix*, 2013).

. Los elementos importantes del cluster:

- NameNode: Sólo hay uno en el cluster. Regula el acceso a los ficheros por parte de los clientes. Mantiene en memoria la metadata del sistema de ficheros y control de los bloques de fichero que tiene cada DataNode.
- DataNode: Son los responsables de leer y escribir las peticiones de los clientes. Los ficheros están formados por bloques, estos se encuentran replicados en diferentes nodos.

El proceso MapReduce

MapReduce es un proceso batch, creado para el proceso distribuido de los datos. Permite de una forma simple, paralelizar trabajo sobre los grandes volúmenes de datos, como combinar web logs con los datos relacionales de una base de datos OLTP, de esta forma ver como los usuarios interactúan con el website.

El modelo de MapReduce simplifica el procesamiento en paralelo, abstrayéndonos de la complejidad que hay en los sistemas distribuidos. Básicamente las funciones Map transforman un conjunto de datos a un número de pares key/value. Cada uno de estos elementos se encontrará ordenado por su clave, y la función reduce es usada para combinar los valores (con la misma clave) en un mismo resultado.

Un programa en MapReduce, se suele conocer como Job, la ejecución de un Job empieza cuando el cliente manda la configuración de Job al JobTracker, esta configuración especifica las funciones Map, Combine (shuttle) y Reduce, además de la entrada y salida de los datos. (*Dataprix*, 2013).

- HDFS: El sistema propio de Hadoop. Está diseñado para la escala de decenas petabytes de almacenamiento y funciona sobre los sistemas de archivos de base.
- Amazon S3. Éste se dirige a clusters almacenados en la infraestructura del servidor bajo demanda Amazon Elastic Compute Cloud. No hay conciencia de racks en este sistema de archivos, porque todo él es remoto.
- CloudStore (previamente llamado Kosmos Distributed File System), el cual es consciente de los racks.
- FTP: éste almacena todos sus datos en un servidor FTP accessible remotamente.

HTTP y HTTPS de solo lectura.

¿Quién usa Hadoop?

En la web de Hadoop podemos encontrar una gran lista de empresas que utilizan Hadoop para gestionar su sistema de archivos. Multinacionales como Adobe, Ebay, Amazon, Facebook o Google trabajan con Hadoop a la hora de mover grandes cantidades de datos y realizar búsquedas. Yahoo! es la compañía que actualmente tiene más dependencia de la herramienta. (Delgado, 2017).

4.4 HDFS. Bases de datos SQL y noSQL.
Hadoop Distributed File System (HDFS)

De acuerdo con Caminero & Grau (2016) Es un sistema de archivos distribuido y tolerante a fallos. Funciona sobre el conjunto de los nodos de un cluster de Hadoop, balanceando la carga de archivos entre las máquinas del cluster, de forma equitativa. Gracias a su naturaleza distribuida, proporciona alta disponibilidad y altas prestaciones que le permiten ser capaz de manejar grandes ficheros.

Para insertar datos en HDFS existen una variedad de formas:

· Copiarlos manualmente utilizando un comando.

· Utilizando la herramienta Flume, que recoge datos de diversas fuentes y los inserta automáticamente.

· Utilizando la herramienta Sqoop, que transfiere datos entre HDFS y bases de datos relacionales.

(Caminero & Grau, 2016)

Escrituras en HDFS

La forma en que HDFS gestiona las escrituras de archivos se explica a continuación:

1. Primero, el fichero de datos se divide en bloques de tamaño fijo, normalmente 64 o 128 MB. Esto se muestra en la Figura.
2. Tras esto, cada bloque se almacena en varios de los nodos del cluster. Esto se muestra en la Figura.

De esta forma, al estar cada bloque de datos replicado en varios nodos del cluster, en caso de fallo de alguno de los nodos, no se pierde información, y el sistema puede seguir funcionando correctamente (exceptuando el decremento de la capacidad del sistema consecuencia del fallo).

Para gestionar HDFS, tenemos un nodo especial en el cluster que se llama NameNode. Esta máquina almacena para cada fichero dónde se almacenan los bloques que lo forman.

(Caminero & Grau, 2016)

Lecturas en HDFS

El proceso de lecturas de ficheros en HDFS es como sigue:

1. En primer lugar, el ordenador del cliente realiza la petición de lectura de un archivo al NameNode (esto se muestra en la Figura 10).

2. Entonces, el NameNode chequea en sus registros qué bloques pertenecen a dicho archivo, así como dónde están almacenados tales bloques, y devuelve esta información al cliente (ver Figura 11)

3. Tras esto, el cliente solicita directamente a los nodos correspondientes que le envíen los bloques de dicho fichero (ver Figura 12). Finalmente, los nodos le envían al cliente los bloques correspondientes directamente, sin pasar por el NameNode.

(Caminero & Grau, 2016)

SQL (Lenguaje de Consulta Estructurado)

El **SQL** (Structure Query Language), es un lenguaje de consulta estructurado establecido claramente como el lenguaje de alto nivel estándar para sistemas de base de datos relacionales. Los responsables de publicar este lenguaje como estándar fueron precisamente los encargados de publicar estándar, la ANSI (Instituto Americano de Normalización) y la ISO (organismo Internacional de Normalización). Es por lo anterior que este lenguaje lo vas a encontrar en cualquiera de los DBMS relacionales que existen en la actualidad, por ejemplo, ORACLE, SYBASES, SQL SERVER por mencionar algunos. (García, A. 2012, p. 3)

El SQL agrupa tres tipos de sentencias con objetivos particulares, en los siguientes lenguajes:

· Lenguaje de Definición de Datos (DDL, Data Definiton Language)

· Lenguaje de Manipulación de Datos (DML, Data Management Language)

· Lenguaje de Control de Datos (DCL, Data Control Language)

(García, A. 2012, p. 3)

A continuación, se describen cada uno de los lenguajes:

Lenguaje de Definición de Datos (DDL, Data Definiton Language)

Grupo de sentencias del SQL que soportan la definición y declaración de los objetos de la base de datos. Objetos tales como: la base de datos misma (DATABASE), las tablas (TABLE), las Vistas (VIEW), los índices (INDEX), los procedimientos almacenados (PROCEDURE), los disparadores (TRIGGER), Reglas (RULE), Dominios (Domain) y Valores por defecto (DEFAULT).

CREATE, ALTER y DROP

Lenguaje de Manipulación de Datos (DML, Data Management Language)

Grupo de sentencias del SQL para manipular los datos que están almacenados en las bases de datos, a nivel de filas (tuplas) y/o columnas (atributos). Ya sea que se requiera que los datos sean modificados, eliminados, consultados o que se agregaren nuevas filas a las tablas del base de datos.

INSERT, UPDATE, DELETE y SELECT.

Lenguaje de Control de Datos (DCL, Data Control Language)

Grupo de sentencias del SQL para controlar las funciones de administración que realiza el DBMS, tales como la atomicidad y seguridad.

COMMIT TRANSACTION, ROLLBACK TRANSACTION, GRANT REVOKE

4.5 Aplicación práctica

Amazon y el Big Data. Una historia de éxito.

¿Cómo ha usado Amazon el Big Data?

Análisis predictivo de las compras.

¿Qué es la filtración colaborativa ítem a ítem?

Mejora de la experiencia de usuario

Aumento de ventas.

¿Cómo ha usado Amazon el Big Data?

El uso de Big Data en Amazon está basado en el machine learning, es decir, la capacidad de los sistemas de aprender a través del análisis de datos masivos que se obtiene de la recogida, almacenamiento y ordenamiento que hace el Big Data. Ejemplo:

Análisis predictivo de las compras. El análisis predictivo de las compras emplea las técnicas de Big Data, como la creación de bases de datos para su posterior análisis mediante aprendizaje automatizado, para crear el sistema de recomendaciones que sugiere a los clientes de Amazon productos relacionados con sus gustos. Este sistema se basa en el análisis de los historiales de compra de los usuarios (recordemos que, para poder comprar en Amazon, necesitamos crear una cuenta de usuario que genera datos cada vez que entramos en la plataforma).

A través del análisis del comportamiento de los usuarios, como qué artículo ha comprado, cuáles ha puesto en el carrito y cuáles ha quitado, qué artículos se han mirado con anterioridad, qué sección suele visitar más, qué ha puesto en su lista de deseos, etc., Amazon puede crear listas personalizas de recomendaciones basadas en una predicción de artículos en los que será más seguro que el cliente acabe picando.

Además, Amazon cuenta con herramientas que permiten a empleados con perfiles muy variados, sin necesidad de tener uno técnico, visualizar datos de forma comprensible para poder aplicarlos en su trabajo.

¿Qué es la filtración colaborativa ítem a ítem?

La filtración colaborativa ítem a ítem es una técnica basada en el Big Data, que emplea Amazon para mostrar a cada usuario desde el momento que entra en la página aquellos productos o artículos que tiene más probabilidades de considerar adquirir. Muestra así productos basados en los propios gustos del consumidor, en lo que ha mirado y comprado con anterioridad, lo que aumenta significativamente las posibilidades de que se acabe produciendo una compra, especialmente si se trata de novedades.

Mejora de la experiencia de usuario.

Este sistema de recomendaciones tiene su máxima expresión en «Mi Amazon.es», una página destinada a mostrar solo recomendaciones a los usuarios; una página que, además, es diferente para usuario de Amazon, puesto que está construida en base al comportamiento de cada usuario en la plataforma. Así como en las recomendaciones que aparecen si hacemos un pequeño scroll hacia abajo al consultar un artículo en particular, apareciendo productos similares, pero también relacionados con el comportamiento previo del usuario. Gracias al Big Data training (o entrenamiento de Big Data) al que se someten los algoritmos que emplea Amazon para analizar los datos recopilados y aprender de ellos, la plataforma es capaz de ofrecer una experiencia completamente personalizada a cada usuario, lo aumenta la fidelización de este, que ve en Amazon no solo una forma rápida y cómoda de comprar, sino también un lugar donde le resulta fácil encontrar aquello que quiere (e incluso a aquello que no sabía que quería).

Aumento de ventas.

El análisis predictivo de compras y la personalización de la experiencia de usuario gracias a ese sistema de recomendaciones y sugerencias ha logrado que Amazon aumente sus ventas año a año, prueba de ello es cómo cada Black Friday y CyberMonday consigue batir récords de venta online.

El hecho de que Amazon te recomiendo productos basados en tus compras y búsquedas anteriores aumenta las posibilidades de que hagas una compra compulsiva al ver un nuevo producto o un producto que no habías visto antes.

Big data en Amazon AWS.

El éxito de Amazon integrando el Big Data en sus procesos de negocio es una realidad.

La forma en la que Amazon ha integrado el Big Data en sus procesos de negocio le han traído un éxito que nadie puede negar y millones de usuario utilizan este enorme comercio electrónico, que empezó como una tienda de libros online, día a día. Y aunque en esta entrada nos hemos centrado más en la parte de ventas, lo cierto es que la recogida y análisis de datos está presente en otras áreas de Amazon, como la logística o atención al cliente, así como en Prime Video, que cuenta con un sistema de recomendaciones similar al de la tienda. Gracias al Big Data, Amazon ha conseguido «conocer perfectamente» a sus clientes y prever su comportamiento a la hora de navegar por su plataforma.

Amazon no esconde completamente el secreto de su éxito e incorpora las funciones Big Data en Amazon Web Services (o AWS, la plataforma de soluciones en la nube de la compañía). Los servicios Big Data de AWS incluyen no solo el almacenamiento y análisis de datos, sino también sus análisis predictivos y aprendizaje automático. AWS cuenta con diferentes herramientas que se pueden contratar para sacarle todo el partido a al Big Data que recopilan la empresa, todo dentro de un entorno en la nube, completamente seguro y escalable, puesto que las cantidades de datos siguen creciendo de manera exponencial.

Amazon Transporte: ¿Por qué su gestión es un gran ejemplo?

1.Centrarse en la experiencia del cliente

El enfoque de Amazon está completamente orientado a cumplir o superar las expectativas de sus clientes, y aunque Amazon se especializa en B2C, ahora los clientes B2B esperan que las respuestas por parte de sus socios de transporte y logística sean similares a la de Amazon. También brindó servicios nunca escuchados, como entrega el mismo día, garantía del producto y devoluciones fáciles, asegurando la satisfacción del cliente y su lealtad a través de la membresía de Amazon Prime.

2. Las torres de control

Para aquellos que no lo saben, una torre de control de logística brinda visibilidad de los productos que no solo están en stock, sino los que están en ruta hacia los clientes. Este amplio seguimiento de mercancías permite a las empresas de logística identificar tendencias y responder a ellas de manera oportuna. Nucleus Research, en un informe de 2016, predijo que Amazon dejaría fuera del negocio

3. Los datos

En Amazon, el enfoque en el cliente no está sujeto a opiniones, están impulsados por los datos. Data Selinger, fundador y director ejecutivo de Rich Relevance, comenzó su carrera en Amazon. Mientras trabajaba directamente con Bezos, se le asignó la tarea de estudiar los datos de Amazon para encontrar nuevas formas de hacer crecer el negocio. Afirma: “Si bien la sabiduría convencional ha sostenido que el servicio al cliente es la salsa secreta de Amazon, la innovación central de Bezos fue colocar los datos en el centro de su cultura corporativa”. En Amazon, todo se mide, no solo la logística. Desde reseñas de productos y rendimiento del sitio web hasta RR.HH. y cuentas, todo se convierte en datos para analizar. Este enfoque ayuda a la empresa a planificar y ejecutar sobre la base del “panorama general”, a diferencia de una empresa de logística que mide solo aquellas métricas que están directamente relacionadas con su negocio. Los conocimientos que se extraen de este Big Data son, naturalmente, mucho más efectivos en la evolución de las políticas de acuerdo con las expectativas del cliente y el sentimiento del mercado.

4. Orientación tecnológica.

Amazon siempre ha estado a la vanguardia de la adopción de tecnología. La organización nunca se ha sentido cómoda con los métodos predominantes de hacer negocios, mientras busca alternativas tecnológicas mejores y más inteligentes. Amazon, siendo la principal empresa de tecnología, me atrevo a decir que debe dedicar tiempo y esfuerzo en descubrir cómo puede aprovechar las tecnologías emergentes para su beneficio. Al adquirir tecnología puedes impulsar a tu organización a una posición estratégicamente importante en un mercado altamente competitivo. Si hay algo que Amazon siempre ha hecho bien, es dirigir las decisiones comerciales basadas en datos, todas sus decisiones se han basado en las tendencias y los conocimientos proporcionados por los números. Si bien no podemos, ni debemos, imitar a Amazon, ciertamente podemos inspirarnos con su enfoque, los datos nos rodean en todo momento, es solo que algunos de nosotros optamos por medirlo, mientras que otros optan por priorizar otros factores al tomar decisiones. ¿Y tú? ¿Crees que las empresas tienen mucho que aprender del enfoque de Amazon transporte? ¿Cómo utilizar los datos para orientar tus decisiones?

Preguntas referentes a la unidad

1.- Una de las mejores herramientas que usa Big Data es:

2.- Python es un lenguaje de programación orientado a objetos como:

3.- Grupo de sentencias de SQL que soportan la definición y declaración de los objetos de la base de datos.

4.- Según los datos de Stack Overflow, Python es el lenguaje de programación de más rápido crecimiento en el mundo. Verdadero o falso

5.-Apache Sparck es un motor de procesamiento de datos de código abierto realmente rápido. Verdadero o Falso

6.-El cloud computing es una nueva tendencia que está ayudando a 3 modelos de Big data, **EXCEPTO**.

7.- Enfoque de Amazon que esta completamente orientado a cumplir o superar expectativas de los clientes.

8.-En la actualidad se identifican 4 tecnologías diferentes para el almacenamiento de grandes volúmenes de informacion, **EXCEPTO**.

9.-Python R no se consideran lenguajes de programación esenciales para la ciencia de datos, lo ideal sería dominar ambos para tener una base de programación completa. Verdadero o Falso

10.- MongoDB es una base de datos orientada a documentos. Verdadero o Flaso.

Respuestas: 1.- Apache Hadoop, 2.- Javascript o C++ 3.- Data Definition Language, 4.- Verdadero, 5.- Verdadero. 6.- Call centers, 7.- Experiencia del cliente, 8.- Memoria USB, 9.- Falso, 10.- Verdadero.

REFERENCIAS

Ali, H. S., Arshad, M. J., & Sumra, I. A. (2019). 7, Vs of Big Data: A Survey.

Ángel M. Rayo. (2016). Tipos de datos en Big Data: clasificación por categoría y por origen. Sitio web: https://netmind.net/es/tipos-de-datos-en-big-data-2/

Argonza, J. S. (2012). Dispositivos para el almacenamiento de grandes volúmenes de información “Big data”. Obtenido de Google schoolar en: Obtenido de Google schoolar en: https://bit.ly/38Bnjxl

Ariel Ortiz Ramírez. (2010). Python como primer lenguaje de programación. 19-05-2010, de Tecnológico de Monterrey, Campus Estado de México Sitio web: https://arielortiz.info/publicaciones/primer_lenguaje_30_jun_2010.pdf

Calles, A. (2022, 3 enero). Amazon Transporte: ¿Por qué su gestión es un gran ejemplo? Drivin. https://blog.driv.in/es/amazon-transporte-por-que-su-gestion-es-un-gran-ejemplo/

Caminero Herráez, A.C & Grau Fernández, L. (2016). *INTRODUCCIÓN AL MANEJO DE DATOS MASIVOS CON HADOOP.* UNED. Recuperado de: https://www.cartagena99.com/recursos/alumnos/apuntes/Practica-SBD-2015-16_v1.pdf

del Conocimiento, I. D. I. (2020, 28 agosto). Las 7 V del Big data: Características más importantes. Instituto de Ingeniería del Conocimiento. https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/

Delgado, D. O. (2017, 27 agosto). *¿Qué es Hadoop? introducción*. OpenWebinars.net. https://openwebinars.net/blog/que-es-hadoop-introduccion/

Equipo de edX. (2021). R vs. Python para la ciencia de datos: Explicación y consejos de aprendizaje. 19-05-2022, de Desconocido Sitio web: https://blog.edx.org/es/r-vs-python-para-la-ciencia-de-datos-explicacion-y-consejos-de-aprendizaje#:~:text=Python%20es%20mucho%20m%C3%A1s%20sencillo,m%C3%A1s%20sencilla%20para%20el%20an%C3%A1lisis.

García, A. (2012). *manual práctico de SQL.* {lwp} Comunidad de programadores. Recuperado de: https://www.lawebdelprogramador.com/cursos/archivos/ManualPracticoSQL.pdf

Introducción a Hadoop y su ecosistema | *Dataprix TI*. (2013, 3 abril). Dataprix. https://www.dataprix.com/es/blog-it/jcasanella/introduccion-hadoop-y-su-ecosistema

Joyanes L. (2013). Big Data: Análisis de grandes volúmenes de datos en organizaciones. México: Alfaomega.

Méndez A. (2012) “Big Data: ¿humo o reto corporativo?” www.penteo.com [Consultado: 18 de Mayo de 2022]. Disponible en Internet: https://bit.ly/3yKEU0r

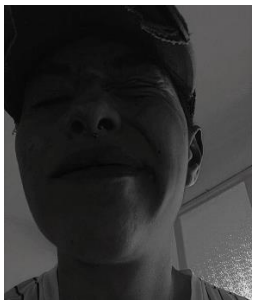
Pérez, M. (2015). *BIG DATA-Técnicas, herramientas y aplicaciones*. Alfaomega Grupo Editor. Obtenido de Google schoolar en: https://bit.ly/3PjkMIV

Poulson, B. (2019, 27 enero). *Una breve introducción a Hadoop - Fundamentos de big data: Técnicas y conceptos*. LinkedIn. https://es.linkedin.com/learning/fundamentos-de-big-data-tecnicas-y-conceptos/una-breve-introduccion-a-hadoop

Redacción España. (2019). 5 herramientas de almacenamiento de datos que se usan en Big Data. ://agenciab12.mx/noticia/5-herramientas-almacenamiento-datos-big-data

ANEXOS

Semblanzas del equipo de trabajo.



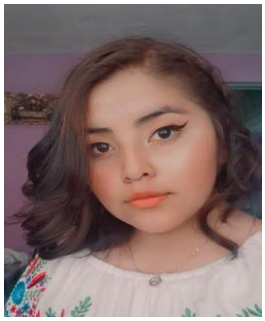
Irving Daniel Aguilar Aguilar actual estudiante en Informática Administrativa Principales intereses: finanzas, inteligencia artificial, Machine Learning, Deep Learning, Redes Neuronales y la Docencia.

Mi nombre es Fátima Cuevas Cruz nacida en el año 2000 originaria del municipio de Jilotepec, estado de México. Con base a esfuerzo y disciplina he logrado poner a nivel mis estudios y por los efectos de la disciplina actualmente me he enfocado a terminar la carrera universitaria.

Mi nombre es Pedro Duran Martínez, originario de la Comunidad San Antonio Enchisi, cuento con el título de Técnico en Informática como formación en El CBT Lic. Mario Colin Sánchez, Atlacomulco, Actualmente cursos la licenciatura en Informática Administrativa en el centro Universitario UAEM Atlacomulco, mis

interese se realizan por ayudar a comprender y descifrar lo que es desconocido por las personas, una persona competente que da lo mejor para destacar en la vida.

Esmeralda Garcia Carmen, originaria del Estado de México actualmente estoy cursando octavo semestre de la licenciatura informática administrativa en el centro universitario UAEM Atlacomulco. Me siento muy motivada de realizar e innovación de contenidos digitales. Pienso que completar mi formación académica adquiriendo conocimientos en contenidos digitales me ayudará mucho a desarrollar mi carrera profesional en un futuro.

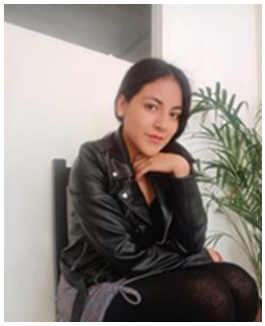


Ana Aletvia Hernández Romero; originaria de la comunidad del Rincón de la Candelaria, Municipio de Atlacomulco, Estado de México, actualmente me encuentro cursando el octavo semestre de la Licenciatura Informática Administrativa en la Universidad Autónoma del Estado de México en el Centro Universitario UAEM Atlacomulco. Permitiéndome una nueva oportunidad con la implementación de este proyecto como apoyo y experiencia a mi persona; así mismo obteniendo habilidades y conocimientos.

Efrain Mateos Casimiro. Posee intereses en la parte de software y seguridad informática y como pasatiempo actividades como lo es el ciclismo. Actual mente estudiante de la Licenciatura en informática Administrativa en el Centro Estudiantil de la Universidad Autónoma del Estado de México.

Juan Ignacio Ortega Sánchez originario de San Francisco Tepeolulco, Temascalcingo, México actualmente estudiante de la licenciatura en informática administrativa cursando el octavo semestre en la Universidad Autónoma del Estado de México, estoy adquiriendo conocimientos sobre las tecnologías de la información, además de como emplear herramientas para el análisis de datos para llegar a una conclusión y búsqueda de soluciones en este caso para la presentación de proyecta BI.

Jose Eduardo Retana Contreras actual estudiante en Informática Administrativa Principales intereses: finanzas, sistemas informáticos y docencia.



Julia Jimena Ruiz Macedonio. Aficionada por la lectura, gusto por los deportes, actualmente cursando la Licenciatura en Informática Administrativa dentro del Centro Universitario UAEM Atlacomulco, acercamiento continuo a la cocina y la naturaleza.

Cristian Segundo Romero, actualmente estudiante de la licenciatura en informática administrativa cursando el octavo semestre en la Universidad Autónoma del Estado de México, obteniendo habilidades sobre las tecnologías de la información, en este caso para la presentación de proyecta BI.

Solís Colín Iván

Nacido en el municipio de Xalatlaco, Estado de México, adicando actualmente en Atlacomulco de Fabela.

FORMACIÓN ACADEMICA

Cursando actualmente la Licenciatura en informática administrativa en el Centro Universitario UAEM Atlacomulco (octavo grado).

AREAS DISCIPLINARES D E DESARROLLO A**CA**D EM**CO**
Según el perfil de egreso de la licenciatura, se tienen aptitudes en:**Gestión** de sistemas de Información Administrativa dentro de una **organización.**

- Planear y programar funciones para el funcionamiento óptimo de los SIA.
- Diseñar SIA para mejorar flujos de información
- Adiestrar al personal en el uso de las TIC's

Administrar proyectos informáticos que optimicen los recursos asignados.

- Determinar proyectos de innovación.
- Crear planes de manejo de recursos.
- Monitorear y controlar avances.

Crea y asesora negocios electrónicos

- Desarrolla ideas de negocios
- Crear un plan de negocios
- Estudia los recursos tecnológicos de la empresa

Audita sistemas de gestión en seguridad de la información

- Identifica áreas de mejora.
- Prioriza acciones de optimización y seguridad
- Establece políticas y procedimientos
- Implementa y evalúa la seguridad en las organizaciones

INTERESES

- Investigación.
- Análisis de datos.
- Desarrollo de software.

PARTICIPACIONES

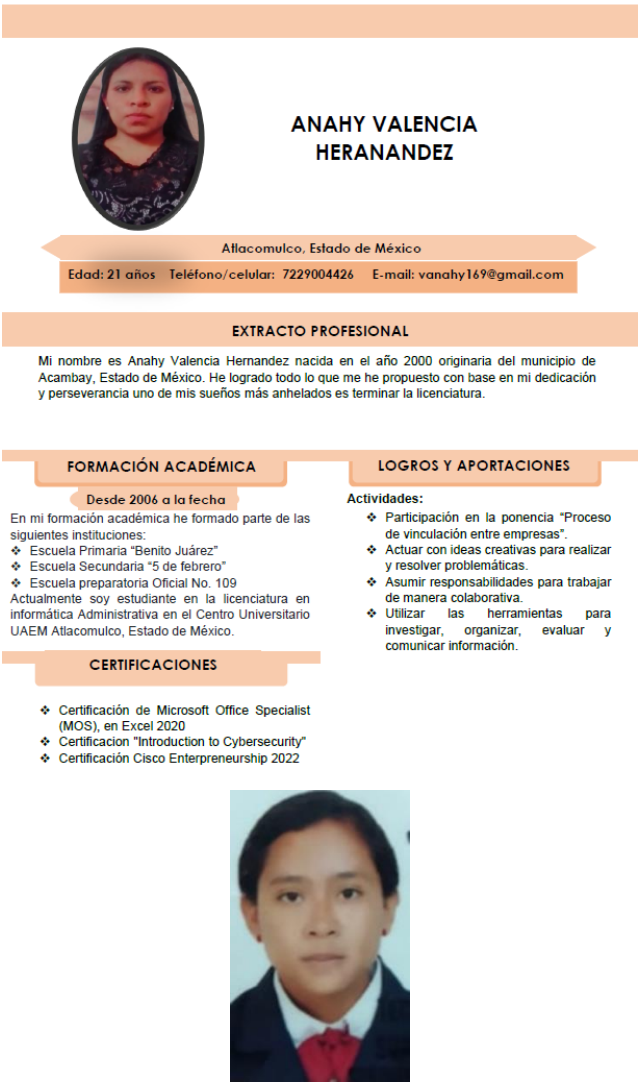
A lo largo de esta permanencia en el centro universitario UAEM Atlacomulco, estancia, se hatenido participación en las siguientes actividades;

- ✓ Servicio Social (Hospital General de Atlacomulco por 7 meses.
- ✓ Obtención de certificación como ciberseguridad por cisco.

Formación Académica.

- ✓ Preescolar
- ✓ Primaria (Josefa Ortiz de Domínguez).
- ✓ Secundaria (EST. No. 99).
- ✓ Preparatoria (COLEPOS) Colegio Nacional de Educación Profesional Técnica).
- ✓ Actualmente cursando la educación superior en el centro universitario UAEMex Atlacomulco en el octavo semestre en la carrera de informática administrativa.

Juan Pablo Ugalde Zaldivar, Atlacomulco México. 11 de abril del 2000. Actualmente estudiante del octavo semestre de licenciatura en informática administrativa. Principales intereses: inteligencia de negocios, gestión de sistemas e investigación. Logros obtenidos certificación en inglés.



ANAHY VALENCIA HERANANDEZ

Atlacomulco, Estado de México

Edad: 21 años Teléfono/celular: 7229004426 E-mail: vanahy169@gmail.com

EXTRACTO PROFESIONAL

Mi nombre es Anahy Valencia Hernandez nacida en el año 2000 originaria del municipio de Acambay, Estado de México. He logrado todo lo que me he propuesto con base en mi dedicación y perseverancia uno de mis sueños más anhelados es terminar la licenciatura.

FORMACIÓN ACADÉMICA

Desde 2006 a la fecha

En mi formación académica he formado parte de las siguientes instituciones:

- Escuela Primaria "Benito Juárez"
- Escuela Secundaria "5 de febrero"
- Escuela preparatoria Oficial No. 109

Actualmente soy estudiante en la licenciatura en informática Administrativa en el Centro Universitario UAEM Atlacomulco, Estado de México.

LOGROS Y APORTACIONES

Actividades:

- Participación en la ponencia "Proceso de vinculación entre empresas".
- Actuar con ideas creativas para realizar y resolver problemáticas.
- Asumir responsabilidades para trabajar de manera colaborativa.
- Utilizar las herramientas para investigar, organizar, evaluar y comunicar información.

CERTIFICACIONES

- Certificación de Microsoft Office Specialist (MOS), en Excel 2020
- Certificacion "Introduction to Cybersecurity"
- Certificación Cisco Enterpreneurship 2022

Nombre: Brenda Vázquez Clemente. Formación académica: Estudiante del CU UAEM Atlacomulco. Interés: Investigación y lectura.

Alfredo Vázquez Ramírez. Enfocado en estadísticas, aproximaciones, proyecciones temporales con Business Intelligence, interesado en análisis prospectivos. Principales intereses o características profesionales del investigador.

YEPEZ MARTÍNEZ DIEGO MAURICIO

Nacido en Toluca, Estado de México

PARTICIPACIONES

A lo largo de la estancia en el Centro Universitario UAEM Atlacomulco, se ha tenido participación en las siguientes actividades;

- ✓ Dos semestres en el programa de mentorías académicas.
- ✓ Dos veces aplicador de exámenes de admisión.
- ✓ Obtención de certificaciones, destacando TOEIC, Excel y ciberseguridad
- ✓ Presentación de proyecto "Análisis de patrones electorales según el PREP 2015"
- ✓ Realización de proyecto (aun en curso) "Auditoria informática bajo el marco de referencia COBIT 4.1 a la Unidad de Tecnologías de información y comunicaciones en el CU UAEM Atlacomulco"

RECONOCIMIENTOS Y DISTINCIONES

- ✓ Certificado de participación por los dos semestres como mentor académico.
- ✓ TOEIC con una puntuación de 634 puntos.
- ✓ Obtener el rol de consejero alumno representante de la Licenciatura en Informática Administrativa
- ✓ Introducción a la ciberseguridad por CISCO
- ✓ Constancias de participación en jornada de ponencias Oracle

FORMACIÓN ACADEMICA

Actualmente cursando la Licenciatura en informática administrativa hasta el octavo grado en el Centro Universitario UAEM Atlacomulco.

AREAS DISCIPLINARES DE DESARROLLO ACADEMICO

Según el perfil de egreso de la licenciatura, se tienen aptitudes en:

Gestión de sistemas de Información Administrativa dentro de una organización.

- Planear y programar funciones para el funcionamiento óptimo de los SIA.
- Diseñar SIA para mejorar flujos de información
- Adiestrar al personal en el uso de las TIC´s

Administrar proyectos informáticos que optimicen los recursos asignados.

- Determinar proyectos de innovación.
- Crear planes de manejo de recursos.
- Monitorear y controlar avances.

Crea y asesora negocios electrónicos

- Desarrolla ideas de negocios
- Crear un plan de negocios
- Estudia los recursos tecnológicos de la empresa

Audita sistemas de gestión en seguridad de la información

- Identifica áreas de mejora.
- Prioriza acciones de optimización y seguridad
- Establece políticas y procedimientos
- Implementa y evalúa la seguridad en las organizaciones

INTERESES

- Auditoría informática
- Docencia
- Mantenimiento preventivo y correctivo de equipos de computo
- Consultor de necesidades de TI