UNIVERSIDAD NACIONAL “SAN LUIS GONZAGA DE ICA”

FACULTAD DE INGENIERIA DE SISTEMAS

PROYECTO DE TESIS

ANALISIS Y USOS DEL BIG DATA APLICADO EN LA UNIVERSIDAD NACIONAL “SAN LUIS GONZAGA” DE ICA: CASO FACULTAD DE INGENIERÍA DE SISTEMAS

BACHILLER: ALMORA CASTRO, NELLIE CATALINA

ICA-PERU

2018

DEDICATORIA

Dedico esta tesis a mis padres Luis Enrique Almora Mendoza y Luzmila Leonor Castro Cárdenas de Almora, ya que gracias a su invaluable apoyo, preocupación e incondicional amor, especialmente su interés constante día tras día, he podido llegar a la culminación de este trabajo, así como lograr mi desarrollo profesional.

Por ese inmenso cariño, amor y dedicación que me han brindado hasta la fecha.

AGRADECIMIENTOS

En primer lugar, a mis padres, por siempre estar presente y ayudar a encaminarme en esta larga jornada. Gracias por estar siempre presente.

A mis padrinos, que siempre confiaron en mi y apoyaron a través de sus consejos y experiencias, a ellos que en todo momento demostraron preocupación por mi formación.

A mi asesor, por su paciencia y comprensión en el desarrollo del tema, a él que supo verter sus conocimientos y capacidad para la culminación de la meta trazada.

A mis docentes, por su sabiduría y consejos, por haberme apoyado y guiado hacia el camino correcto durante mi tránsito por las aulas universitarias con mucha dedicación.

A mis compañeros, por haber compartido innumerables y buenos momentos, por haber compartido tantas e invaluables experiencias en nuestra convivencia.

RESUMEN

Los Big Data son un conjunto de datos grandes y complejos, tanto estructurados como no-estructurados, que necesitan distintos instrumentos para su almacenamiento, así como su búsqueda y visualización. Pero, ¿qué más hay detrás de este término? ¿qué otros usos se le puede atribuir? Generalmente todo lo que concierne a este es reducido debido a la falta de información o incluso a la falta de interés en uno de los temas y tecnologías menos aprovechadas de los últimos años. La recolección de datos bastamente grande, compleja y variable, y que no puede ser gestionada con instrumentos tradicionales es solo parte de las características que definen al Big Data como tal.

Cuando hablamos de Big Data, hablamos de cientos de datos que no pueden ser analizados tradicionalmente, como lo sería a través de bases de datos relacionales, porque requerirían de un software adecuada para dichas funciones, ya que los sistemas operativos de un ordenador están ideados para trabajar con datos almacenados en local, entonces, ¿qué pasa cuando los datos almacenados son tan grandes y existe una gran cantidad de estos?

Tener conocimiento de que hacer en cada una de estas situaciones llevó al estudio de muchísimas personas en el campo del manejo de los datos e hizo posible la creación de arquitecturas y tecnologías que permiten el estudio, así como la aplicación de datos que no estaban siendo utilizados y a su vez siendo desperdiciados.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

INTRODUCCION

Desde la aparición de la internet, el mundo se ha acelerado y con esta aceleración, se ha ido generando mucha información. Como es de conocimiento la información es un conjunto de datos, pero estos datos crecen día a día, cada hora, minuto y segundo que transcurren en nuestras vidas. Como tal las diversas organizaciones sean pública o privadas están viendo que mucha información se crea en sus espacios, pero muy poca de esta información es aprovechada.

En el mundo una organización cualquiera sea su rubro u orientación, genera datos, como consecuencia de que las personas dentro y fuera de ella la están generando de una u otra manera, directa o indirectamente; el auge de hoy de la “redes sociales”, ha incrementado más aún la información con otros tipos de datos que hoy se generan, ya no solo son los datos puramente conocidos como textos, e imágenes, sino se ha ido generando otros tipos de datos, como videos, datos geoespaciales, etc.

Esta situación que está pasando en el mundo con la generación exponencial de los datos ha motivado a que muchas de las empresas relacionadas con la tecnología hayan orientado sus esfuerzo a qué hacer con eso datos y como aprovecharlos, en tal sentido nace el concepto de Big Data, o el tratamiento de grandes volúmenes de datos, basado en los puntos precedente, la presente investigación pretende hacer un acercamiento a esta nueva necesidad de las organizaciones y tener un mayor conocimiento sobre Big Data y como se está desarrollando en el mundo.

Nellie Almora

CAPITULO I: EL PROBLEMA Y OBJETIVOS

1.1. El Problema de Investigación

Es de conocimiento de todos los que estudiamos Ingeniería de Sistemas, que lo más importante en una organización, son los datos y la información que se genera de ella y más aún en una denominada como “la era del conocimiento”, en donde es importante que las organizaciones conozcan y que tengan un pleno conocimiento de que está pasando en su organización, desde el punto de vista de los datos. De aquellos datos que están circulando dentro de su organización, pero que muy poco se sabe de estos, en tal sentido el tratamiento de dichos datos para convertirlo en información y esta información igualmente gestionarla adecuadamente para poder convertirla en conocimiento, es el reto de las organizaciones de hoy.

De otro lado los datos ya no solo son los datos convencionales. Sino que se está generando datos de mayor tamaño como audio, video, mapas geoespaciales, etc. que obligan a encontrar nuevas formas de poderlo tratar. Las personas hoy en día emplean las redes sociales para generar datos que las organizaciones no están aprovechando, los empleados, nuestros clientes, proveedores, visitantes están demandando “algo” que está escondido detrás de esos datos y que es importante su tratamiento.

Nuestra universidad Nacional San Luis Gonzaga, alberga cerca de 18,000 estudiantes de pregrado y cerca de 1,000 estudiantes de posgrado según el Director General de la Oficina de Registro Matricula y Estadística (OGMRE), igualmente se tienen cerca de 1,000 personal administrativo. Vale decir tenemos solo en nuestra universidad cerca de 20,000 personas que están generando datos sobre la universidad, los alumnos generando cantidad de información de matrícula, notas etc. Los docentes generando cantidad de información sobre sus temas de interés, al igual que los trabajadores que están generando mucha información de su interés que la universidad desconoce.

Como se desprende de un análisis muy general solo sobre nuestra universidad, existen gran cantidad de datos que se están generando, pero que dichos datos están ahí ocultos, sin ser aprovechados y nuestra universidad tienen muchas carencias en cuanto a información que pueda ser aprovechada por los tomadores de decisiones en cada nivel. A nivel de la alta dirección o en las facultades a los decanos y directivos que podrían hacer uso de dichos datos, previo tratamiento.

A pesar que es de nuestro conocimiento que hay muchos datos que se están generando en la universidad, nos preguntamos ¿estamos preparados para poder hacer uso de esos datos?, ¿Los estudiantes de la facultad de ingeniería de sistemas están preparados, o cuentan con el conocimiento necesario?

1.1.1. Situación Problemática

Basado en la información de los problemas descritos en el punto 1.1. se desprende y es de imaginarse que en nuestra provincia se está generando millones de datos que cada segundo están generando nuestra población sobre sus temas de interés, y viene aquí las interrogantes ¿Qué está demandando nuestra población?, ¿Qué necesidades son requeridas por la población estudiantil de la UNICA?, ¿Qué están demandando los docentes de nuestra Universidad?, ¿Cuáles son sus necesidades del personal administrativo?, en fin se abre un abanico de interrogantes sobre esto; pero además hay poco conocimiento sobre el Big Data, ¿cómo se aplica?, ¿qué herramientas son necesarias?, ¿qué arquitectura de es requerida para poder aprovechar estos datos?, etc., como tal con las interrogantes anteriormente planteadas, nos permitimos formular la siguiente interrogante.

1.1.2. Formulación del problema

¿Cuál es el grado de conocimiento del Big Data, y que herramientas están asociadas a su aplicación, en los estudiantes de la FIS-UNICA?

1.1.3. Delimitación del problema

El problema está delimitado a la facultad de ingeniería de sistemas, la misma que como carrera profesional, se desea conocer el conocimiento que se tiene sobre esta Herramienta de Gestión de Grandes Volúmenes de datos.

1.2. Justificación e Importancia

El Big Data para los que hemos estudiado Ingeniería de Sistemas, es un concepto nuevo y no muy claro, pero que además nos hemos preguntado siempre y como se tratan grandes volúmenes de datos, nos imaginamos solamente la cantidad de información que procesos Facebook, como procesa la información Twitter, donde se encontraran sus servidores, que características tienen esos servidores o "Granja de Servidores". Para los que estamos ligados a la Facultad de Ingeniería de Sistemas de la UNICA, con la expectativa muy grande creada en nosotros sobre la implementación de la HPC (High Performance Computer) o "supra computadora", con ella ¿se podrá aplicar Big Data?.

De lo anterior, si para los que hemos estudiado Ingeniería de Sistemas nos es aún vago el concepto, ya que este solo es aplicado por grande organizaciones a las que no tenemos acceso, imaginemos a las personas comunes, las personas de a pie como siempre se mencionan, que saben que hay mucha información en las redes, o los Gerentes, Administradores de las organizaciones en nuestra ciudad que les es más vago aún el concepto, a pesar que ellos mismos están generando mucha información. He ahí la importancia sobre su conocimiento e investigación.

1.3. Objetivo

El objetivo que se persigue con esta investigación es entender con mayor amplitud al Big Data como nueva tecnología para grandes volúmenes de datos, que herramientas están relacionadas con él, así como el alcance de los datos para que sean aprovechados correctamente.

De este objetivo se desprende algunos objetivos específicos, como a) descubrir el conocimiento que se tiene en los estudiantes de la FIS-UNICA sobre el Big Data, b) motivar al conocimiento del Big Data en los estudiantes de la FIS-UNICA, c) profundizar en los conocimientos sobre el Big Data en la FIS-UNICA.

CAPITULO II: MARCO TEORICO

2.1.Antecedentes

José Antonio Rojas García en el año 2016, propone un modelo de negocios basado en Big Data que facilite la integración de los datos de las personas naturales y de soporte a las políticas de e-government en el Perú, apoyado en una empresa de logística integral en su tesis de pregrado en la Universidad Peruana de Ciencias Aplicadas UPC, Lima-Perú. Presenta una "Propuesta de Negocio" para una Empresa de Logística Ligera que busca aprovechar sus habilidades operativas a fin de desarrollar un nuevo servicio para sus clientes basado en Big Data. La propuesta se encuentra alienada a la prioridad y capacidad competitiva de la Empresa de Logística Ligera considerando sus procesos operativos actuales y su plan estratégico actual, a la vez se plantea el tratar de aprovechar su capacidad instalada actual, minimizando su capacidad ociosa actual que se ha generado a consecuencia de atender a sus clientes actuales. La propuesta que se ha planteado tiene como principal objetivo atender las necesidades de los consumidores futuros de la Empresa de Logística Ligera, así como que esta pueda servir de soporte para varias de las políticas implementadas en la actualidad para el desarrollo de un e-government en el Perú. La "Propuesta de Negocio" presentada se soporta en la construcción de un modelo de Big Data confiable y sostenible en el tiempo acorde a la normatividad vigente para manejo de datos personales y que pueda ser parte la misma de los procesos que permitan mejorar la planificación de servicios que ofrece el Estado Peruano a los ciudadanos. El principal objetivo es actualizar los datos personales de las personas naturales en un lapso máximo de 48 horas en todas aquellas instituciones que se encuentren afiliadas al modelo de negocio propuesto, para lograr este objetivo se proponen procesos de validación de los datos recabados, adicionalmente como se propone la transmisión de dicha información de forma estructurada y enriquecida en tiempo real a fin de que esta pueda ser utilizada en un modelo dinámico de información interconectada a lo largo del territorio nacional. Finalmente se podrá generar beneficios adicionales que podrán ser complemento y

soporte de algunas de las políticas del Estado Peruano para los próximos años como lo es la planificación de los servicios para la sociedad mediante la identificación real del número de ciudadanos en cada ámbito geográfico, así como la reducción de los costos de las organizaciones y de los ciudadanos para actualizar los datos de contactabilidad, generando de esta manera la transformación de la sociedad actual en una en una nueva "Sociedad Digital".

Emilcy Juliana Hernández Leal en el año 2016 propone la aplicación de técnicas de análisis de datos y administración de Big Data ambientales en su tesis de maestría para la Universidad Nacional de Colombia. El crecimiento en el volumen de datos generados por diferentes sistemas y mediciones de actividades cotidianas en la sociedad es un factor que influencia directamente en la necesidad de modificar, optimizar y concebir métodos y modelos de almacenamiento y tratamiento de datos que suplan las falencias que presentan las bases de datos y los procesos de KDD tradicionales. Big Data es un enfoque que incluye diferentes tecnologías asociadas al almacenamiento, análisis y visualización de grandes volúmenes de datos provenientes de diferentes fuentes y que se presenta como una solución ante los problemas de tratamiento de datos que no son cubiertos por las soluciones tradicionales; cabe anotar que cuándo se hace referencia a grandes volúmenes de datos, no hay un consenso entre los autores respecto a una cantidad a considerar como grande, en parte puede depender del dominio de los datos.

Por otra parte, el monitoreo de condiciones ambientales como las climáticas, meteorológicas e hidrometeorológicas constituyen una fuente de datos que puede aumentar de manera exponencial, en la medida en que se hagan mediciones de estos fenómenos en diferentes periodos de tiempo, ubicaciones espaciales y estrategias de captura.

Teniendo en cuenta los planteamientos anteriores, se pretende por medio de esta tesis, la concepción de un modelo para la administración y análisis de datos ambientales con el uso de algunas tecnologías Big Data, que permita facilitar el tratamiento de estos datos, su almacenamiento, aplicar diferentes tipos de análisis y extraer información relevante de apoyo a la toma de decisiones y en general a la comprensión de los datos propios del dominio.

María Fernanda Laverde Salazar en el año 2015 propone el diseño de un curso teórico y práctico sobre Big Data en su tesis de maestría para la Universidad de Chile. La veloz expansión del uso de la tecnología, genera un conjunto de desafíos en cuanto al manejo y análisis de grandes cantidades de datos que se generan a una gran velocidad, ya que se debe lidiar con situaciones vinculadas tanto con los datos, el software & hardware y además la relaciones entre clientes y proveedores de servicios. El Big Data es una etapa en la era digital, y no representa un concepto aislado, ya que para su correcto aprovechamiento es necesario establecer una integración con los métodos de análisis de datos que permitirán sacar provecho a la información recolectada. La posibilidad de tomar decisiones y luego llevar a cabo acciones útiles a través de los resultados obtenidos, mediante herramientas de análisis de datos, es lo que constituye el núcleo del Big Data Analytics. Tal como lo expresa Michael Minelli, coautor del libro Big Data: "Big Analytics: Big Data" no es sólo un proceso para almacenar enormes cantidades de data en un data warehouse, sino también es la habilidad de tomar mejores decisiones y tomar acciones útiles en el momento preciso. El trabajo de grado que se desarrolla a continuación corresponde al diseño e implementación de un curso teórico y práctico sobre Big Data. Dicho curso está orientado a alumnos de pregrado de la Universidad de Chile, y se basa en un diseño curricular siguiendo una metodología docente específica la cual se estructura en bloques de planificación y desarrollo. El programa se divide en módulos de

aprendizaje definidos mediante temas y objetivos, y tiene una duración de cuarenta (40) horas en total entre clases teóricas (20 horas divididas en 10 clases) y prácticas (20 horas divididas en 5 laboratorios). El objetivo general del curso, es integrar conocimientos relacionados con la forma de almacenar, administrar y aprovechar mediante herramientas específicas, el incremento sustancial del volumen de datos que se manejan diariamente, e inclusive cada segundo, en las empresas de tecnología y comunicación de las cuales, en su mayoría, día a día somos los principales generadores de data.

Fernando Manso en el año 2015 propone el análisis de modelos de negocios basados en Big Data para dispositivos móviles en su tesis de maestría para la Universidad de San Andrés. En este trabajo se realiza una descripción de la problemática del modelo de negocios actual de los operadores de telecomunicaciones móviles y la industria. Se introducen los conceptos básicos a nivel de las soluciones de Big Data, como éstas pueden satisfacer necesidades que el Business Intelligence tradicional no puede y se presentan casos de uso de operadores móviles líderes. Finalmente, se analizan los modelos de negocio de internet y las formas de monetización gratuita como base de la propuesta de un nuevo modelo de negocio basado en la explotación de datos de los operadores móviles.

Miguel Lostaunau Fuentes en el 2015 discute los problemas de uso de datos sobre el crimen en los informes del estado en su tesis de maestría para la UPC. Desde hace algunos años el tema de la seguridad ciudadana se ha ido situando como un problema público que atrae las miradas de diversos actores de la sociedad y del ámbito político en el Perú, esto en razón de que se trata de un problema que vulnera funciones del Estado y derechos del ciudadano, es por ello que se requiere una acción efectiva por parte del Gobierno para reducir los niveles de criminalidad que afectan a la población.

Esta necesidad que expresa la ciudadanía responde a una de las principales labores que desempeñan tanto el Estado como la Policía Nacional, no obstante cuándo la literatura

sobre las funciones y capacidades del Estado denoten que se trata de una tarea fundamental y presente desde el inicio de la instauración de esta forma de gobierno, es esencial reconocer que los fenómenos delictivos van variando y se requiere una adaptación constante para poder enfrentarlos.

Es por ello que una de las bases que sustenta la labor policial es el recojo sistemático de datos y estadísticas que coadyuven un correcto diagnóstico y posterior elaboración de una política o estrategia de seguridad con resultados concretos en relación a la reducción del crimen, sin embargo, el caso peruano es un ejemplo de un inadecuado manejo de las estadísticas sobre criminalidad.

Este hecho ocasiona que las distintas instituciones involucradas en el combate del crimen (como la Policía Nacional, el Ministerio Público, Poder Judicial, INPE, el INEI, etc.) mantengan datos distintos que se traducen en la adopción de acciones diferenciadas y desligadas a la hora de combatir el crimen, de manera que no se apunta a objetivos en común ni a una focalización especial en sectores especialmente críticos de nuestro país.

Dicha investigación tiene como objetivo mostrar la problemática de la data estadística y su manejo en los informes del Estado y mostrar la existencia de una "cifra negra", además se busca mostrar la caracterización de la estadística criminal basada solo en función a las denuncias realizadas.

Este desorden ha dado lugar a que distintos informes del Estado sobre la criminalidad y la seguridad ciudadana tengan falencias y no reflejen la realidad de lo que afirman. Esta afirmación nos lleva a plantear la hipótesis si esta no sería una de las causas del fracaso de las diversas políticas públicas sobre Seguridad Ciudadana. La revisión de los diagramas de flujos del procesamiento de la data estadística demuestra que las instituciones siguen procedimientos erróneos y que al ser distintos estarían ocasionando un mucho mayor esfuerzo que termina siendo inútil, así como un uso de recursos para el

logro de un objetivo que podría centralizarse y servir como un verdadero respaldo para la toma de decisiones en materia de políticas de seguridad.

Fabián Andrés Guerrero López y Jorge Eduardo Rodriguez Pinilla en el año 2013 expusieron el diseño y desarrollo de una guía para la implementación de un ambiente Big Data en la Universidad Católica de Colombia en su tesis de pregrado para la Universidad Católica de Colombia. En dicha investigación ambos autores concluyeron que el Big Data es una nueva tendencia para el manejo de grandes volúmenes de información, utilizado principalmente por grandes empresas, pero gracias a las nuevas tecnologías y su fácil acceso podrá ser utilizado por cualquier empresa o institución que desee vincularse al nuevo proceso que se puede lograr en la gestión de la información.

La estructura de un ambiente Big Data ayuda a mejorar la manipulación de los datos, optimizando la gestión de la información respecto a tiempo y costo, logrando obtener mejores resultados en las estadísticas para una buena toma de decisiones.

La creación de un ambiente Big Data se debe realizar dentro de un cluster, el cual permita integrar todas las aplicaciones que se van a utilizar, como en este caso Hadoop, en el cual se almacena la información y las aplicaciones corren dentro del mismo nodo, evitando conflictos durante la ejecución.

Es importante resaltar que existen muchas maneras para transformar el mismo modelo relacional al modelo basado en columnas, ya que se pueden tomar distintos caminos para la unión de los datos, esto depende de la información que se desee encontrar o saber. Para obtener una adecuada transformación se deben tener en cuenta las llaves primarias, las cuales se convertirán posteriormente en row key, permitirán integrar toda la información dentro de una misma columna, mejorando la manipulación que se darán a los datos.

David López García en el año 2013 presentó el análisis de las posibilidades de uso de Big Data en las organizaciones en su tesis de maestría para la Universidad de Cantabria. En estos tiempos que corren con el conocimiento como preocupación principal, y en una sociedad en cual los clientes y las empresas están cambiando, dichos grupos cada vez generan e intentan procesar más y más datos, cantidades que para muchos son imposibles de imaginar. Para lograr adquirir y analizar tanta información surge el término "Big Data". Un término joven que presenta confusión respecto a su alcance.

En este trabajo se tratará de aclarar en qué consiste, su alcance, como lo utilizan las empresas y en qué situación se encuentra. Además, también se abarcará otros términos relacionados con Big Data, como pueden ser la minería de datos, el Cloud Computing o el Data Warehouse. Igualmente, también se aclarará porqué surge Big Data, de dónde procede y por qué para muchos tecnólogos sugiere un cambio de etapa en el mundo de las Tecnologías de la Información.

2.2. Bases Teóricas

2.2.1. Big Data

La información que se genera cada segundo en el mundo es exponencial, en relación a la información que podemos consumir, hay grandes volúmenes de datos que están viajando por la red, información de las personas que día a día están haciendo algo en un mundo online. Los datos son más voluminosos, los clientes de las empresas igualmente están generando muchos datos que las empresas no están aprovechando de manera efectiva, talvez porque aún hay un desconocimiento en nuestro país tanto en la región, como también en nuestra universidad, porque Big Data es aún un concepto relativamente nuevo del que no se están aprovechando sus ventajas. Heredia y Nieto (2017) en su propuesta de una metodología para la adopción del Big Data, menciona que el "(..) volumen creciente de datos, además de variado, puede contener información

valiosa para otras personas y organizaciones", sin embargo "en la actualidad las técnicas y tecnologías tradicionales ya no son suficientes para este fin", por lo cual "han surgido nuevas ciencias cuyo objetivo son los datos". Sin duda esta orientación sobre la producción de grandes volúmenes de datos abre nuevas oportunidades para las empresas, Calvo y OSal (2017) recalcan que "recabar información precisa, detallada y continuamente actualizada de sus impactos económicos, sociales y medioambientales". Por otro lado Alegre (2017) en su artículo de cómo sacar partido del análisis de datos Big Data, menciona que "Google, Amazon, Netflix son empresas pioneras en el uso de Big Data, además de que muchas empresa igualmente están aprovechando el potencial de los datos para mejorar sus procesos y crear nuevas prestaciones", igualmente la autora determina que los datos que manejan son una "variedad de datos no estructurados pudiendo ser estos numéricos, textuales, como comentarios en Facebook, Twitter o en el blog de la empresa, o multimedia, como fotografías, canciones, etc."

Como se desprende de estos conceptos iniciales, que los datos están presentes en todo ámbito sobre los que giran las personas, en tal sentido Ureña, Tenesaca y Mora (2017) en su investigación sobre Análisis y procesamiento de datos académicos de una institución superior con herramientas de Big Data, en la que "hoy en día estamos en la era de generar, a un ritmo exponencial gran cantidad de datos en tiempo real ya sean estos estructurados, semi-estructurados y no estructurados, provenientes de diversos orígenes como las redes sociales, tablets, celulares, sensores, entre otros (…), el problema en la actualidad no es la generación exageradas de datos, sino la forma en que se almacenan, la velocidad en que se analizan y que resultados se obtienen". Según los autores "cada día en el mundo se generan más de 2.5 exabyte de datos. Esto equivale a 1,000.000 de terabyte", pero además el autor menciona que "el 90% de los datos guardados en la actualidad, han sido creados en los dos últimos años".

Victoria Gómez (2016) en “Bigdata Sales Leader, IBM Eruropa, España”, la tendencia de las aplicaciones móviles, se estima que para el año 2020 se va generar aproximadamente 4 Exabyte de datos, de los cuales el 90% se van a generar en los dos últimos años, datos que deben ser aprovechados por las empresas. Como tal las empresas están cambiando sus modelos de negocios los mismos que giran alrededor de los datos, cada vez se crean modelos de negocio alrededor de los datos; la nube va tomando mayor peso, y menor corso en tal sentido más gente la está adoptando.

José Carlos Baquero (2016) en “Ingeniería de Software y Desarrollo GMV, España”, el Big Data es importante porque cada vez el valor de los datos se vuelve más importante, y estos están asequibles, por lo que el Big Data está para quedarse; en tal sentido la economía está cambiando, cambiando alrededor de los datos, cuyo valor se vuelve más, como tal podemos tener un modelo de negocio ya que sin tener un solo taxi, puede ser la mayor flota del mundo.

Richard Bengamin (2016) en “Director BI & BIGDATA, Telefónica, España”, el Big Data es un revolución, que muchas veces no viene de la empresa tradicional.

Mar Castaño (2016), Directora Data Science de Territorio creativo de España comenta que la transformación digital está obligando a las distintas compañías a crear nuevos modelos de gestión, operación en nuevos contextos y ambientes; por un lado el ambiente desde el punto de vista para el cliente, ya que no puede mantenerse ajeno a la empresa y tiene que interactuar por medio del nuevo ecosistema digital, plataforma como es el recorrido del cliente, que tipo de servicio desea contratar, valores como este cliente se está informando ¿cuándo? ¿dónde? ¿en qué momento?, puntos de contacto, cómo se relacionan. En qué medida los volúmenes de los productos y servicios son importantes, su experiencia, como consumo.

Igualmente, la autora detalla en cuanto al ámbito del negocio, negocio digital, nuevos canales de ventas de sus productos, e-comerce, determinan algunos productos que pueden evolucionar en un ecosistema digital. Las compañías tienen que evolucionar, las personas van a tener que ser conocedoras del modelo digital y abordar ese proceso, todo lo que tiene que ver con datos y el Ecosistema Big Data, acompañado de una revolución de datos que se deben recabar en todos los niveles y tomar decisiones dirigidas por datos.

2.2.2. Importancia del Big Data

La transformación digital, y la enorme cantidad de datos que se genera, abre nuevas oportunidades para el aprovechamiento de dichos datos y es aquí donde el Big Data ofrece una oportunidad relevante; Mar Castaño (2016) de territorio creativo, en relación al Big Data como solución aporta más valor, al proceso de expansión del cliente, recibiendo información de las redes sociales, datos de la compañía; todo lo que tiene que ver con el comportamiento se analiza de distintas fuentes que son muy importantes en el recorrido con el cliente donde se focaliza más los esfuerzos, al final el dato es el dato y tiene aplicación en todos los niveles; por este lado nuevas innovaciones de productos y servicios se están generando, por lo que es muy importante recabar información de tendencias, lo que está haciendo la competencia y de nuestras propias capacidades. La geolocalización también está al día, el open data como origen de datos complementa a los análisis que actualmente se realizan.

2.2.3. Usos del Big Data

Para entender los usos del Big Data, hay que conocer no solo el tamaño de los datos, sino de otras características en la que los expertos la sintetizan en las tres “V”; como tal Inés Alegre, Miguel Ángel Ariño y Miguel Ángel Canela, profesores de Análisis de Decisiones IESE, nos detallan estos conceptos, en relación al Volumen, en la que estamos habituados a hablar de megabytes o de gigabytes, en Big Data podemos hablar

de órdenes de magnitud muy superiores, que pueden llegar a los Terabytes (1000 gigas) o incluso a los exabytes (1000 millones de gigabytes), se trata pues, de grandes volúmenes de datos difíciles de manejar por un ordenador personal con el software convencional, por otro lado otra de la características es la Velocidad, por la que en algunas aplicaciones de Big Data, los datos se generan de forma continua y se han de procesar en tiempo real, o dentro de una ventana de tiempo que se va reduciendo progresivamente. Es lo que llamamos streaming data.

2.2.4. Herramientas para Big Data

Son muchas las herramientas que son aplicables para el concepto de Big Data, algunas de las que han sido estudiadas como Hadoop, HDFS, No SQL, MapReduce, MongoDB, Pig, HIVE y HBase herramientas que trabajan juntas para lograr el objetivo de extraer valor de los datos (Ureña, Tenesaca y Mora, 2017).

Vidal, Bustamante, Lapo y Nuñez (2018), nos detallan sobre algunas herramientas para realizar trabajos con Big Data, una de estas es MapReduce, el cual es un enfoque de computación para trabajar con grandes volúmenes de datos en un entorno distribuido con altos niveles de abstracción y con el uso ordenado de funciones Map y Reduce, la primera de ellas para el mapeo o identificación de datos relevantes y la segunda para resumir datos y resultados finales. Por otro lado los autores indican que MapREduce es una metodología de programación dada a conocer por Google (Guller, 2015; Dean y Ghemawat, 2004) para la computación distribuida sobre grandes cantidades de datos o Big Data, "MapReduce ha tenido una masiva adopción por medio de Hadoop, una implementación libre de código abierto" (Apache Hadoop, 2016; White, 2015; Lin y Dyer, 2013).

Felipe Ortega, profesor de la Universidad Rey Juan Carlos de España, en su Debate sobre "Tendencias Futuras en Análisis de Datos de Big Data", en la que se analiza herramientas orientadas hacia Big Data, con el panel Victoria Gómez, José Carlos

Baquero y Richard Bengamin, analizaron las diversas tecnologías Big Data, en tal sentido Victoria comenta que a pesar de que no están aún muy desarrolladas, se están trabajando en Tecnologías Cognitivas con orientaciones a sectores como sanidad y educación, la idea es democratizar más el Big Data para que sea fácilmente accesible. Con la tecnología cognitiva, la idea es que se sea capaz de aprender de los datos entendiendo el lenguaje natural; por otro lado, José Carlos Baquero menciona que se hacen necesarios entornos colaborativos de datos, para aunar esfuerzos más ágiles, siendo en este caso aún un entorno joven y que requiere mayor estandarización con plataformas Open Data. Además, Richard Begamin, explica sobre el Data Wrangling, cuya finalidad es escoger los datos crudos y luego llevarlos a que tengan calidad, que se entiendan bien, que tenga metadata, por lo que se requiere tecnologías que lo automaticen.

2.2.5. Fuentes de datos para Big Data

Las fuentes de datos son variadas, así como su almacenamiento; como es de conocimiento se tiene base de datos relacionales, base de datos no relacionales (Data Warehouse), como también almacenamiento No SQL; en tal sentido los datos que se manejan para Big Data son datos de diversas naturalezas estructurados, semi-estructurados y no estructurados, dichos datos provienen de diversas fuentes de datos de las redes sociales, y cualquier dispositivo que almacene datos.

2.2.6. Otras Tecnologías Asociadas al Big Data

Sin duda hay otras tecnologías asociadas al Big data, la inteligencia de negocios, que nace con el concepto de Dataware House que manejan información de gran parte de la empresa, asociadas a este igualmente se tiene Data Mining. Además, tecnologías en la "nube" o Cloud Computing, y las herramientas que han hecho que los datos exploten a niveles inimaginables como lo serían las redes sociales, fuentes bastante importantes en el tema de Big Data.

2.2.7. Cloud Computing

Este es un término que ha ido creciendo mucho en el ámbito de la tecnología, desde sus inicios hubo mucho escepticismo en cuánto al resguardo de información. Este concepto tiene como importancia la facilidad y economía que tiene su implementación; la "nube" como es conocida son servicios que se ofrecen en todos los niveles, servicios de infraestructura, aplicaciones, etc, para el portal de IBM "Los productos de Cloud pueden maximizar los beneficios y optimizar las cargas de trabajo. Aun así, IBM Cloud es mucho más que un simple cloud. El catálogo de servidores de IBM Cloud contiene potentes opciones que satisfacen los requisitos y presupuestos de muchas aplicaciones. Así, puede hacer crecer su negocio en la medida que lo necesite de forma rápida, segura y fiable".

2.2.8. Redes sociales

Las redes sociales hoy en día es un concepto muy conocido y amplio en la población, para la mayoría de personas el término Facebook, Twitter, Instagram, etc, es de su conocimiento y eso constante; estas redes son una de las fuentes más ricas de grandes volúmenes de datos y que son aprovechadas por las empresas con la aplicación de Big Data. Según el portal de Web empresa 2.0 no sólo se consolida Facebook como la gran red social en el primer puesto ya rozando los dos mil millones de usuarios, sino que además sus recientes adquisiciones, Whatsapp e Instagram no dejan de crecer. WhatsApp, que nació en 2009, ha sobrepasado a YouTube ocupando la segunda posición en número de usuarios. Y para rematar el despegue total de Facebook Messenger.

2.2.9. Data Mining

Para el profesor Felipe Ortega de la Universidad Rey Juan Carlos de España, la minería de datos no parte directamente del análisis de datos, y que en su manejo en cuanto a las tareas a realizar existen varias alternativas, en una encuesta entre el año 2002 al 2014 la metodología líder para la minería de datos es la CRISP-DM (Cross Industry Standard

for Data Mining), la cual describe el autor del curso Data Mining, Tendencias en análisis y visualización de datos, la misma que se compone de los siguientes pasos:

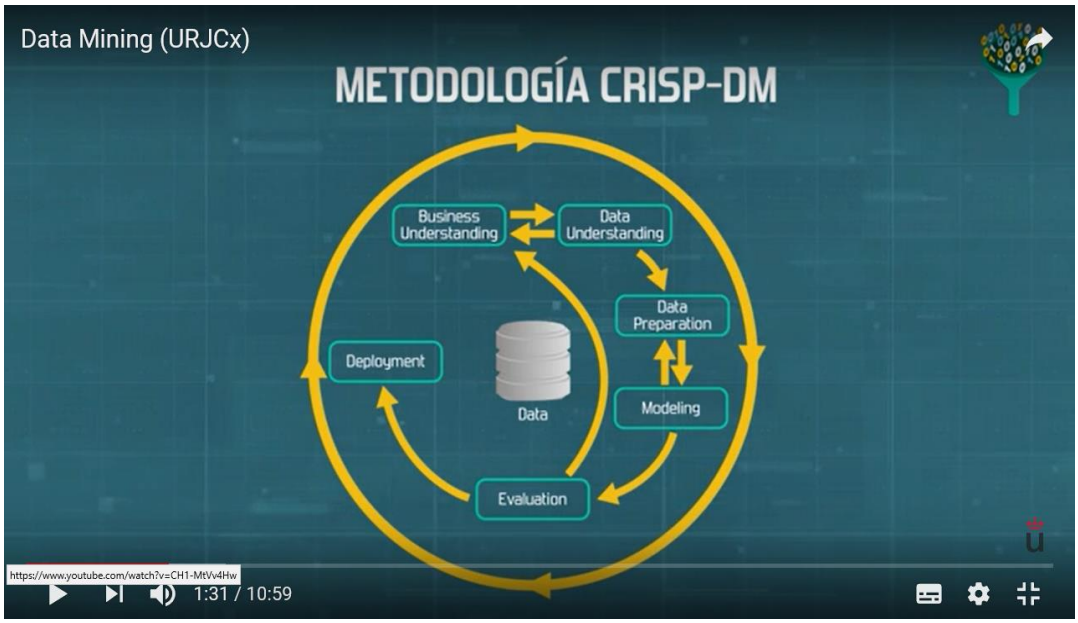


Figura N° 1: Metodología para Data Mining

1° Comprensión del negocio: La parte más importante en un proceso de Data Mining, es que requiere conocer, comprender, identificar los procesos de negocio, especialmente los parámetros y aspectos clave que influyen en el proceso de mejora de dicho proceso de negocio, solo así se puede orientar al objetivo e identificar la pregunta o preguntas de interés que queremos responder mediante el análisis de datos; sin embargo sino se tienen dichas preguntas se pueden realizar análisis exploratorios de datos, que nos puedan guiar a una búsqueda preliminar.

El proceso de Data Mining debe reflejar en todo momento una buena comprensión del proceso de negocio.

2° Comprensión de los datos: Teniendo definido la pregunta o preguntas de interés, se debe conseguir una colección de datos que creemos nos permitan obtener respuestas adecuadas; el análisis debe familiarizarse con los datos conociendo detalles

precisos como su origen significado dentro del proceso de negocio, la codificación, su naturaleza cualitativa o cuantitativa, en que unidades están expresadas. Utilizan técnicas de exploración de datos. El resultado es un “Diccionario de Datos”.

3° Preparación de datos: Parte de la metodología que abarca aproximadamente el 85% del tiempo, el analista prepara de forma metódica los datos y para ello realiza las siguientes tareas:

a. Limpieza de datos, para eliminar posibles problemas de codificación y datos incorrectos.
b. Formato de los datos, para poder leerlos correctamente, pasar los datos de las diversas fuentes a un formato que facilite su análisis posterior.
c. Imputación de datos, para datos faltantes por medios estadísticos.

4° Modelado: Aplicar técnicas y modelos matemáticos y estadísticos para diseñar modelos que permitan contestar nuestras preguntas de interés a partir de la información que nos proporcionan los datos. Herramientas de modelos estadísticos, análisis mediante sistemas de Base de Datos y Data Warehouse; técnicas de inteligencia artificial y técnicas de Machine Learning son utilizadas.

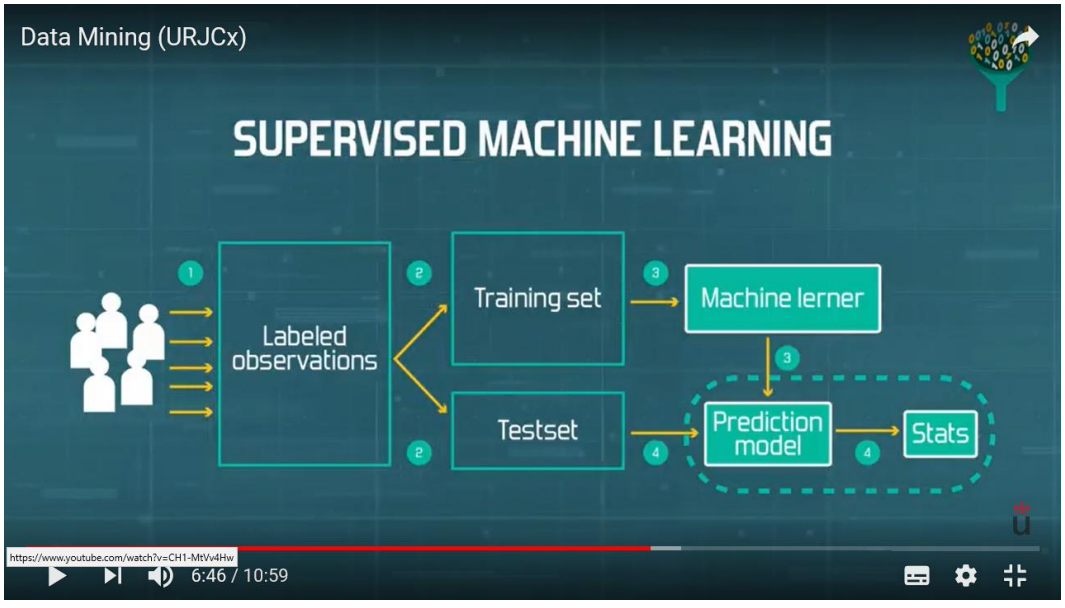


Figura N° 2: Máquina de Aprendizaje

5° Evaluación del modelo: Para ello dividir el modelo anterior en dos o más subconjuntos; un subconjunto se puede utilizar para construir el modelo y los restantes para validar los resultados y generalizar el modelo mediante la validación cruzada en K iteraciones.

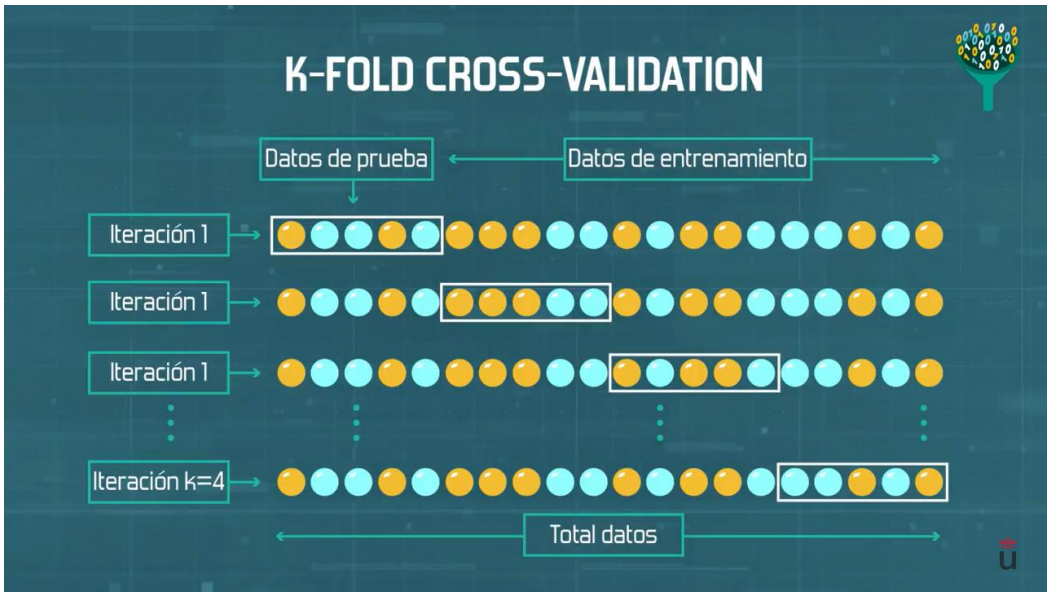


Figura N° 3: Modelo K-Fold Cross Validation

6° Despliegue en producción: el modelo de lleva a producción sobre datos reales

2.3. Marco Conceptual

2.3.1. Big Data

El portal de la empresa DELL EMC se menciona sobre Big Data: “Debido a los avances en la tecnología, la definición de Big Data ha cambiado con el transcurso de los años, aún así algo permanece sin cambios: la cantidad de datos crece de forma continua a una velocidad sumamente rápida, todos los tipos de datos, en cualquier formato, que se usan para obtener información valiosa y generar valor se consideran Big Data”.

2.3.2. Cloud Computing

Para el portal argentino de La Nación (19 de julio de 2011), es el nombre dado al procesamiento y almacenamiento masivo de datos en servidores que alojen la información del usuario.

2.3.3. Red Social

Es un término originado en la comunicación, estas se refieren al conjunto de grupos, comunidades y organizaciones vinculados unos a otros a través de relaciones sociales. Esto no ha sido más que el resultado de la convergencia de los medios, la economía política de los mismos y el desarrollo de tecnologías, teniendo como objetivo la interacción de dos o más canales.

2.3.4. Data Mining

Hay tantos datos y una gran cantidad de decisiones que tomar, las organizaciones de todo el mundo se están enfrentando a este dilema: los datos están creciendo, pero ¿y su capacidad para tomar decisiones de acuerdo con esos enormes volúmenes de datos? ¿También están creciendo?. La analítica predictiva ayuda a evaluar lo que sucederá en el futuro, mientras que la minería de datos (Data Mining) busca los patrones ocultos en los datos que pueden utilizarse para predecir el comportamiento futuro. Las empresas, los científicos y los gobiernos han utilizado este enfoque por años para transformar los datos en conocimientos proactivos (SAS, The Power to Know).

CAPITULO III: METODOLOGIA DE LA INVESTIGACION

3.1. Tipo de Investigación, Nivel y Diseño de Investigación

La investigación, será del tipo transversal, por la cual se recopilará la información sobre Big Data y todo lo que se ha estudiado sobre este para analizarlo y presentar los conocimientos que se obtengan de él.

Siendo la información que se conoce sobre el tema, además de no poseer un conocimiento amplio sobre el mismo, la investigación que se va a realizar está en un nivel de investigación exploratoria, Hernández, Fernández y Baptista (2010) "(…) se realizan cuando el objetivo es examinar un tema o problema de investigación poco estudiado, del cual se tienen muchas dudas o no se ha abordado antes, (…)".

Para nuestro diseño basado en que su nivel es exploratorio se tendrá para el mismo un diseño no experimental, Hernández, Fernández y Baptista (2010) "no se genera ninguna situación, sino que se observan situaciones ya existentes, no provocadas intencionalmente en la investigación por quien la realiza".

3.2. Población y Muestra

La población para obtener información sobre el conocimiento de término Big Data, estará conformada por todos los estudiantes de la facultad de Ingeniería de Sistemas de la UNICA, a su vez, de esta población se toma como población objetivo a los estudiantes que han llevado las asignaturas de "Teoría y Diseño de Base de Datos" e "Implementación de Base de Datos" (ciclo 2018-I).

De esta población se extraerá una muestra con la que se aplicarán los instrumentos, debido al alcance de esta investigación de nivel exploratoria, se aplica un muestreo del tipo no probabilístico o dirigido, en la cual la unidad de análisis son estudiantes de Ingeniería de Sistemas que poseen conocimiento sobre Base de Datos, para ello se seleccionarán a estudiantes de un aula que haya cursado dichas asignaturas.

3.3. Técnicas de Recolección de Datos

Se procederá a realizar una entrevista a estudiantes que hayan llevado las asignaturas mencionadas anteriormente, con previa revisión de los sílabos sobre el tema de estudio señalado.

3.4. Instrumentos de Recolección de Datos

Los instrumentos de recolección de datos estarán basados en una guía de entrevista, para recoger la opinión sobre conocimiento del tema en los estudiantes y las fichas documentales para poder recopilar la información de los documentos asociados.

3.5. Técnicas de Análisis e Interpretación de Resultados

Las técnicas de análisis e interpretación de los resultados, se basa en el esquema de la figura, y en la que los datos son recolectados con los instrumentos elegidos para la investigación, siendo la entrevista una de la técnicas más relevantes del estudio, se tienen que organizar adecuadamente los datos que pueden estar en formato de audio o videos, las mismas deben ser transcritas y, para hacer una análisis más profundo de dichas entrevistas, con dicha información se elaborarán las bitácoras de análisis para documentar paso a paso el proceso analítico. De este proceso analítico se irán obteniendo las ideas generales y cuáles son las orientaciones de esas ideas, todo este proceso culmina con la generación de las explicaciones y teorías que se desprenden de la investigación (Hernández, Fernández y Baptista, 2010).

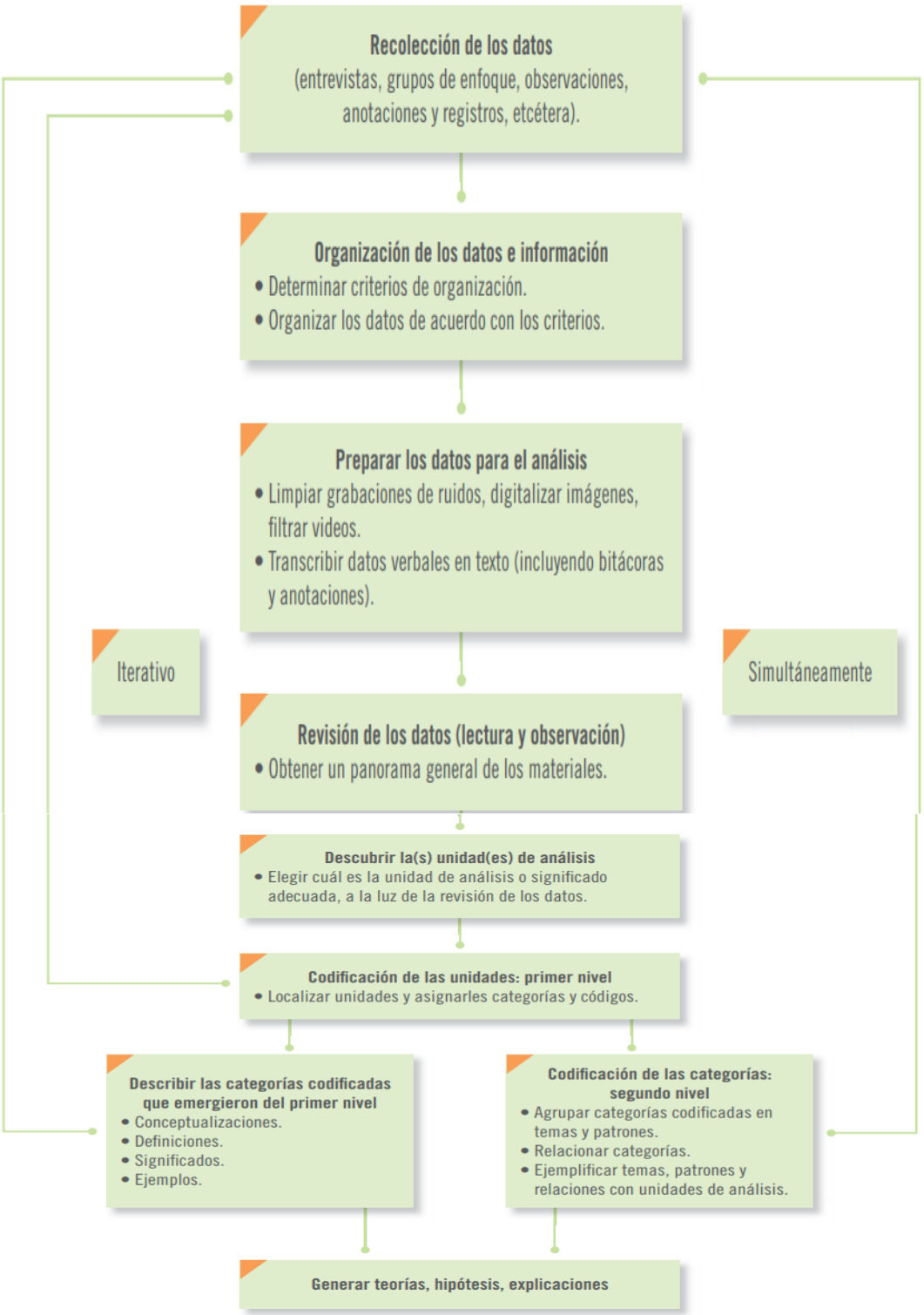


Figura N° 4: Proceso de Análisis Fundamentado en los Datos Cualitativos

Fuente: Hernández, Fernández y Baptista (2010, p. 445)

3.6. Análisis de Big Data

La revolución de los datos masivos abre nuevas oportunidades laborales que deben ser aprovechados por nuestros colegas, para que nos mantengamos a la vanguardia de las mejores instituciones del país y del mundo. Recabar información precisa, detallada y que este constantemente actualizada abre un sinnúmero de oportunidades para que las organizaciones puedan enfrentar un mundo cada día más competitivo, y nosotros tenemos que proveerles a los tomadores de decisiones dicha información.

Ureña, Tenesaca y Mora (2017) aseguran que estamos en la era de generar, a un ritmo exponencial gran cantidad de datos en tiempo real ya sea estos estructurados, semi-estructurados y no estructuradas, proveniente de diversos orígenes como las redes sociales, tablets, celulares, sensores, entre otros, de allí surge el término "Big Data", que nos brinda la capacidad de generar, almacenar y analizar un volumen grande de información para descubrir los diversos comportamientos de los clientes y mejorar la toma de decisiones por parte de los altos ejecutivos de la empresa. De todo esto, el problema en la actualidad no es la generación exagerada de datos (cada día en el mundo se generan más de 2.5 exabytes de datos, equivalente a 1.000.000 de terabytes), sino la forma en que se almacenan, la velocidad en que se analizan y que resultados se obtienen.

Tabla N° 1:
Unidades de Almacenamiento de Big Data

Unidades de almacenamiento		
Unidad	*Symbolo*	*Potencia*
Gigabyte	GB	10^9 bytes
Terabyte	TB	10^{12} bytes
Petabyte	PB	10^{15} bytes
Exabyte	EB	10^{18} bytes
Zettabyte	ZB	10^{21} bytes
Yottabyte	YB	10^{24} bytes
Xerabyte	XB	10^{27} bytes

(Ureña, Tenesaca y Mora, 2017)

Inés Alegre en la revista de negocios del IEEM (agosto 2017), en su artículo cómo sacar partido del análisis de datos Big data, establece que hoy en día son muchas las empresas que están aprovechando el potencial de los datos para mejorar sus procesos y crear nuevas prestaciones; esto supone sin duda que el conocimiento es el resultado del procesamiento de dichos datos, y orienta a las organizaciones a poder desarrollar nuevos conocimientos, nuevos productos, nuevos procesos que es el principio fundamental de la gestión de conocimiento.

Big Data es el nuevo concepto que actualmente la mayoría de empresas buscan implementar en sus negocios con el fin de aprovechar al máximo el gran volumen de información que tienen almacenados en sus bases de datos, buscando obtener un mayor beneficio de sus recursos.

3.6.1 Características del Big Data

De un lado tenemos que, en el Big data la cuestión no es únicamente el tamaño, sino otras características, que los expertos sintetizan en "las tres V", en tal sentido consideran al Volumen, la Velocidad y la Variedad (Alegre, Ariño y Canela, agosto 2017). Estas características se describen tal que:

Volumen: Estamos habituados a hablar de megabytes o de gigabytes, una giga son 1000 megas, mientras que en Big Data, podemos hablar de órdenes de magnitud muy superiores, que pueden llegar a los terabytes (1000 gigas) o incluso a los exabytes (1000 millones de gigabytes). Hablamos pues, de grandes volúmenes de datos que serían difíciles de manejar en un ordenador personal y con el software convencional.

Velocidad: En algunas aplicaciones de Big Data, los datos se generan de forma continua y se han de procesar en tiempo real, o dentro de una ventana de tiempo que se va reduciendo progresivamente. Es lo que llamamos "Streaming Data".

Variedad: A menudo, hablamos de datos estructurados pero al hablar de bases de datos, nos viene a la cabeza la imagen de una colección de tablas, con filas y columnas, en su mayor parte llenas de números. En Big Data además, manejamos datos no estructurados, que pueden ser numéricos; textuales, como comentarios en Facebook, Twitter o en el blog de la empresa; o multimedia, como fotografías, canciones, etc. Pueden provenir de fuentes muy diversas y hallarse en diferentes formatos (CSV, JPEG, PDF, etc.).

La tecnología para trabajar con Big Data empezó a desarrollarse en empresas como Google, Amazon y Netflix, pero su expansión a organizaciones de todos los tipos y tamaños está siendo imparable, tanto para la mejora de los procesos existentes como para crear nuevas prestaciones.

Otros estudios como el de Tenesaca, Ureña, Mora, y Segarra (2017), señalan que el Big Data está basado en cinco características que dirigen su trabajo, conocidas como "las 5 Vs", entre las cuales cuatro se conocen como principales y una como adicional, estas son: Valor, Velocidad, Veracidad, Volumen y Variedad.

Figure 1. Dimensiones de Big Data

Figura N° 5: Dimensiones (características) de Big data

Sin embargo, encontramos otros estudios como el de Doug Laney (citado en Heredia, Nieto, 2017), en donde se establece que las características clave de los datos. El análisis de los autores incluye los casos de los sectores en los que el uso del Big Data está más extendido como el bancario y el de las compañías telefónicas. Los bancos usan datos de sus clientes que captan a través de los pagos con tarjeta y movimientos de sus cuentas

para ofrecerles uno u otro producto. Las compañías telefónicas emplean los datos del uso del teléfono para predecir qué clientes tienen una mayor probabilidad de cambiar de compañía y así planificar una campaña de marketing a medida para lograr retenerlos.
De otro lado se conoce por la investigación que la cadena de establecimientos Macy's ha logrado, en los tres últimos años, un incremento de ventas del 10% gracias al uso inteligente del Big Data en sus operaciones de logística y pricing. Macy's recoge datos en tiempo real de las ventas y el stock disponible en cada una de sus tiendas y adapta los precios y las ofertas de manera dinámica en más de 73 millones de artículos.

Es interesante ver como casos como el de una conocida cadena de comida rápida controla la cantidad de personas que están esperando para ser atendidas. Si la cola es larga, anuncia en sus pantallas productos que son rápidos de cocinar y servir, si no hay cola, anuncia productos que aunque son más lentos de cocinar, tienen un margen de beneficios más alto.

Finalmente, el estudio nos permite conocer que estrategias de precio, de servicio al cliente, de benchmarking, de retención de clientes y marketing son solo algunos ejemplos de los numerosos y variados usos con impacto real en los resultados de negocio que tiene el Big Data. En la mayoría de los casos, no hemos de capturar nuevos datos, sino sacar partido de los que ya tenemos.

A lo largo de lo que se viene desarrollando en relación al Big Data, se dan una serie de metodologías que se utilizan para el mejor tratamiento de esa gran cantidad de datos que hay que procesar.

En el estudio de Heredia y Nieto (2017) presentan una propuesta de una metodología para la adopción del Big Data, y que en los siguientes apartados se irá desmenuzando, es de suponer que volumen de datos, creciente y variado, puede contener información valiosa para otras personas y organizaciones, en la actualidad las técnicas y

tecnologías tradicionales ya no son suficientes para este fin, ya que han surgido nuevas ciencias cuyo objetivo son los datos, el Big Data o datos masivos es el concepto que se impone al hablar de esta gran y creciente cantidad de datos. En el estudio de IDC (International Data Corporation) de 2013, se pronostica que las tecnologías de Big Data y el mercado de los servicios asociados habrá crecido a una tasa seis veces más grande de todo el mercado.

En esta investigación lo que se busca es explorar las diversas metodologías y marcos de trabajo existentes para la gestión y utilización de Big Data en las organizaciones, así como identificar sus características y componentes, como tal la investigación se fundamenta en algunos marcos teóricos como el Data Science y Big data, Data Science es el estudio sistemático de los datos, que implica su organización, propiedades, su análisis y su poder de inferencia; la existencia de Data Science se justifica en el hecho de que los datos, su objeto de estudio, son cada vez más heterogéneos y no estructurados, de tipos cada vez más complejos, como texto, imágenes, videos, entre otros. Data Science brinda un marco de referencia para la extracción de información útil de los datos.

El crecimiento en cantidad y diversidad de los datos ha llevado a que los conjuntos de datos sean tan grandes que se dificulta su manejo por medio de herramientas convencionales, a partir de ello, se han desarrollado nuevos métodos de data science y nuevas aplicaciones, como el análisis predictivo, y se denomina DPB a la confluencia de los conceptos Data Science, Predictive Analytics y Big Data. También influye en la "revolución" del Big Data el aumento significativo en la potencia de los computadores, la ubicuidad de procesamiento y almacenamiento, el incremento en la producción de contenidos digitales y la movilidad.

El analista de datos Doug Laney considera tres características claves de los datos hoy en día, es decir, del Big Data: Volumen, Velocidad y Variedad (Tres V's). Volumen se refiere a gran cantidad de datos, velocidad por la rapidez con que llegan los datos y la rapidez con que pierden vigencia, y variedad porque hay muchos tipos de datos diferentes y cada vez más complejos.

3.6.2. Arquitectura de Big Data

Según la ANSI/IEEE 1471/2000 define a la arquitectura como "Organización fundamental de un sistema, incorporado en sus componentes, la relación entre ellos así como su entorno y los principios que gobiernan su diseño y evolución". Una arquitectura de información provee un marco de trabajo para tratar de forma consistente e integrada con la tecnología para garantizar información confiable que un negocio requiere, una buena práctica para la organización que desea implementar Big Data es adoptar un enfoque de arquitectura empresarial que permita transformar estas iniciativas para mantener la alineación con el negocio y maximizar el retorno de la inversión que en ella se hace.

Se tiene igualmente que la arquitectura del Big Data está basada en capas, las mismas que están determinadas por:

- ✓ Infraestructura: Componentes de hardware y soporte para bases de datos (estructuradas o no) y de Big Data (Tecnologías basadas en Hadoop, por ejemplo).
- ✓ Data: Obtención, depuración, integración y transformación de datos provenientes de múltiples fuentes.
- ✓ Big data y Data Warehouse: Almacenamiento especializado de los datos integrados, transformados y sumarizados, aptos para someterlos a procesos de análisis de diversos tipos.
- ✓ Analíticas de Big Data: Componentes para el análisis de datos.

- ✓ Arquitectura del negocio: Capa de integración entre el sistema de Big Data y los usuarios finales.

Las investigaciones de los autores plantean a algunas dimensiones que son claves para un proyecto de Big Data.

Tabla N° 2: Dimensiones para Proyecto Big data

AREA	DIMENSIÓN
Contexto del negocio	Intención del negocio: Cómo se utilizarán los datos para mejorar el negocio
	Uso de los datos: Qué procesos del negocio se benefician
	Propiedad de los datos: Necesidad de apropiarse de los datos
Visión de la Arquitectura	Alimentación de datos: Características de captura de datos y respuesta del sistema
	Almacenamiento de datos: Tecnologías de almacenamiento apropiadas para el reservorio de datos
	Procesamiento de datos: Estrategia práctica para las aplicaciones basadas en big data
	Desempeño: Cómo maximizar la velocidad de las consultas, las transformaciones de datos y el modelamiento analítico
	Latencia: Cómo minimizar la latencia entre los componentes operacionales claves
	Análisis y descubrimiento: Donde se requiere llevar a cabo el análisis de datos
	Seguridad: Donde se necesita asegurar los datos
Estado actual del Big Data en la organización	Experiencia con datos no estructurados: Se está haciendo algún tipo de procesamiento a los datos no estructurados.
	Consistencia: Se utilizan prácticas estandarizadas de calidad y gobernanza de datos
	Experiencia con tecnologías y herramientas de big data: Nivel de conocimiento y aplicación de ellas
	Habilidades en análisis de datos: Personal científico de datos y analistas familiarizados con técnicas y herramientas avanzadas para análisis de datos.
Estado futuro del Big Data en la organización	Mejores prácticas: Los mejores recursos para guiar la decisión de construir el estado futuro.
	Tipos de datos: Cantidad de transformación requerida para los datos no estructurados en los reservorios de datos.
	Fuentes de datos: Frecuencia de cambio de las fuentes o estructuras de datos
	Calidad de datos: En qué momento aplicar las transformaciones
Mapa de ruta	Prueba de concepto
	Adquisición de habilidades en tecnologías y herramientas
	Adquisición de habilidades en análisis de datos
Gobernanza	Fuentes de datos en la nube: Garantizar la confiabilidad de estas fuentes de datos.
	Calidad de datos: Depuración y enriquecimiento de datos no estructurados y frecuencia de revisión y actualización de las estructuras de datos.
	Políticas de seguridad: Adaptación de las políticas de seguridad a los nuevos requerimientos de big data.

3.6.3. Modelado de Big Data

El modelo de Big Data es una capa abstracta que se utiliza para gestionar los datos almacenados en dispositivos físicos; provee una forma visual de manejar los recursos de datos y permite crear una arquitectura de datos que permita reducir costos computacionales y reusar dichos datos. La capa de modelo de Big Data se encuentra entre la capa física (lugar en el que están almacenados los datos) y la capa de aplicaciones (dónde se hace uso de los datos).

La tendencia actual para muchas empresas de tecnología informática es utilizar bases de datos NoSQL (Not Only SQL), la misma que cuenta con la siguiente clasificación:

- ✓ Clave-Valor (Key-Value): Almacena una clave o identificador y el objeto (de cualquier tipo, simple o complejo) o valor asociado a ella. No requiere un esquema de almacenamiento fijo.
- ✓ Documentales: Permiten un tipo de almacenamiento más complejo que los clave-valor, el uso de índices secundarios y objetos en varios niveles. Soporta datos sin esquema y semi-estructurados.
- ✓ Grafos (Graph Store): La base de datos se representa con un grafo, dirigido y etiquetado, donde los nodos hoja representan datos y los nodos internos representan la conexión entre ellos. Se utilizan cuando es tanto o más importante mostrar la interconectividad que los datos en sí.
- ✓ Bases de datos en columna (Column-oriented data base): Los datos se almacenan en orden de columnas, no en orden de filas como en las bases de datos tradicionales.

3.6.4. Metodología

Han surgido productos y metodologías y marcos de trabajo (frameworks) para Big Data, tanto de código abierto como comerciales. La tecnología que soporta Big Data evoluciona rápidamente y toma tiempo su maduración; a diferencia de otras tecnologías que ayudan a resolver problemas, ésta ayuda a encontrar los problemas. El diseño de sistemas de análisis de Big Data requiere atender ciertos principios, se requieren arquitecturas de alto nivel y frameworks apropiados.

- ✓ Las aplicaciones de Big data deben soportar una variedad de métodos analíticos.
- ✓ No existen soluciones que se ajusten a todos los problemas.
- ✓ El análisis de datos debe estar donde están los datos (en Big Data, el almacenamiento de datos debe ser distribuido).
- ✓ El procesamiento de análisis debe llevarse a cabo en memoria.
- ✓ La coordinación entre las unidades de almacenamiento y las unidades de datos es necesaria para que el sistema sea eficiente y con alta tolerancia a las fallas.

Se propone que una empresa comience con proyectos de Big Data de pequeña escala, para entender la tecnología y las áreas del negocio que pueden beneficiarse; donde se consigna la estrategia de negocio que se pretende abordar con Big Data y debe contener los siguientes puntos: estrategias de negocio (Business Strategy), iniciativas de negocio (Business Initiatives), resultados esperados (Outcomes) y factores críticos de éxito, tareas (Tasks) y fuentes de datos (Data sources).

Geerdink (2013), presenta una arquitectura de referencia, Esta arquitectura de referencia tiene componentes opcionales, los cuales están agrupados en tres capas:

- ✓ Capa de negocios (Business Layer),
- ✓ Capa de Aplicación (Application Layer) y
- ✓ Capa de Tecnología (Technology Layer).

De otro lado Tekiner, Keane (2013), framework que provee las bases para el desarrollo y la gestión de aplicaciones de Big Data, enfocándose en los datos y en el ecosistema de Big data, consiste de tres fases o etapas: Identificación e integración de los datos provenientes de múltiples fuentes: Análisis de datos y modelamiento, Organización de datos e Interpretación de estos mismos.

Igualmente, Vanauer, Hellingrath (2015), establece una metodología diseñada para que las organizaciones estructuren la introducción del Big data a sus procesos, la cual consta de tres etapas generales: Desarrollar ideas para el uso de Big Data, evaluar estas ideas con respecto a su valor potencial, así como los cambios necesarios en la arquitectura empresarial de las organizaciones; e implementarlas coherentemente en el negocio. Algunas literaturas sobre temáticas de Big Data se centran mayormente en infraestructura y analíticas, subestimando el desarrollo de software para este tipo de aplicaciones.

Los autores Heredia y Nieto (2017) en esta investigación proponen una metodología, la misma que se ha desarrollado y que tiene como resultado el siguiente modelo con sus fases las que se analizan a continuación:

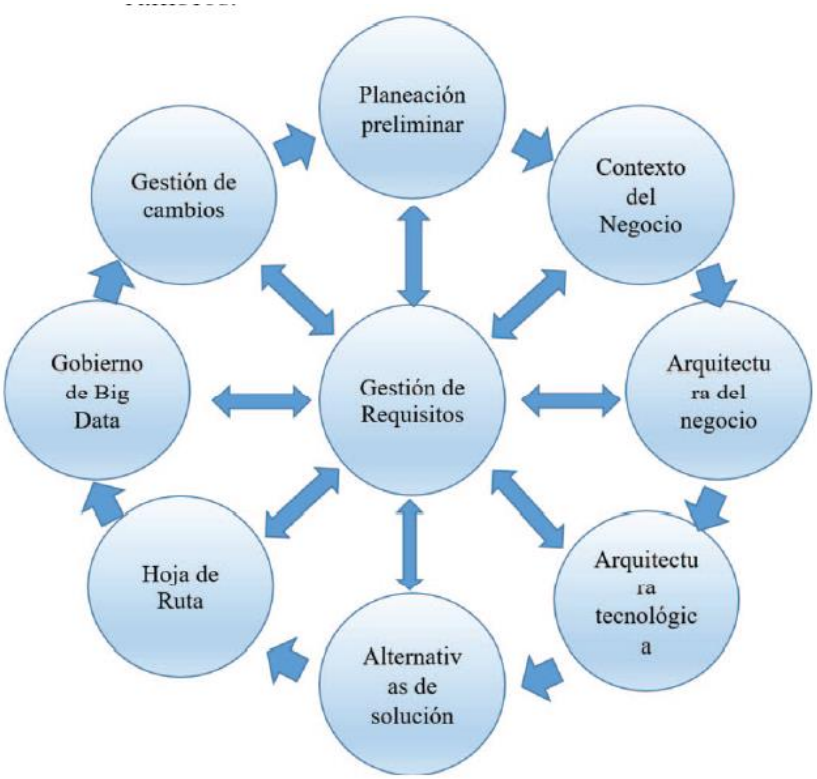


Figura N° 6: Fases de la metodología propuesta

1. Gestión de Requisitos: Identificación y especificación de los requisitos que deben ser implementados en una aplicación basada en Big Data. Es la fase central del modelo, dado que el proceso total es cíclico, en cada nueva iteración se deben analizar los requisitos: Productos, historias de usuario, requerimientos funcionales y requerimientos no funcionales.
2. Planeación Preliminar: Identificación de necesidades, beneficios y capacidades para adoptar Big Data. Productos: Diagnóstico organizacional, alcance del proyecto de Big Data, identificación de activos de datos.
3. Contexto del Negocio: Definición de los procesos del negocio y la estrategia de Big Data apropiada para ellos. Productos: Estrategia del negocio, identificación y especificación de procesos del negocio, especificación de uso de datos para apoyar la estrategia del negocio.

4. Arquitectura del Negocio: Modelamiento del negocio. Producto: Arquitectura del negocio.
5. Arquitectura Tecnológica de Big Data: Identificación de las necesidades tecnológicas de Big Data y diseño de los componentes de la arquitectura tecnológica. Productos: Especificación de tipo y frecuencia de registro de datos, diseño de los repositorios de datos, estrategia de integración de datos, especificación de requerimientos de analíticas, identificación de riesgos sobre los datos.
6. Alternativas de solución: Se identifican y analizan las diversas alternativas de solución. Productos: Descripción de alternativas de solución, evaluación de las alternativas propuestas.
7. Hoja de Ruta: Plan de acción definitivo para la implementación de Big Data. Producto: Hoja de ruta.
8. Gobierno de Big Data: Se definen las políticas de gestión y gobierno de Big Data. Producto: Lineamientos de gobierno de Big Data.
9. Gestión de cambios: Define el proceso de gestión de cambios de cualquiera de los componentes del sistema de Big Data. Producto: Proceso de gestión de cambios.

3.7. Herramienta y tecnología para Big Data

3.7.1. Hadoop

Hadoop es un proyecto de código abierto para aplicaciones confiables, escalables y distribuidas que permite el procesamiento de grandes volúmenes de datos en clúster de servidores, diseñado para extender a un sistema de servidor único a miles de máquinas con un alto grado de tolerancia a fallas. Hadoop es un framework de código abierto desarrollado en el lenguaje de programación Java que mediante el empleo de modelos de

programación simple, permite almacenar y procesar gran cantidad de datos en un entorno distribuido de clústers de ordenadores.

Las características más importantes que posee Hadoop son:

i) Está diseñado para ejecutarse en grupos grandes de hardware conocidos como clúster robustos,

ii) Es considerado como un software muy robusto, debido a que frente a un fallo del hardware puede superar este tipo de problemas,

iii) Es escalable, lo que significa que puede añadirse nodos al clúster con facilidad y,

iv) Permite a los usuarios escribir código con eficiencia.

Igualmente, el estudio de Ureña, Tenesaca y Mora (2017), establece que Hadoop es un framework que de manera transparente provee fiabilidad, es accesible, escalable, robusto, tolerante a fallos y distribuido siendo capaz de administrar cualquier tipo y volumen de datos (Alcivar y Espinoza, 2017). Incluye: MapReduce (motor de cálculo offline), HDFS (sistema de ficheros distribuídos) y HBase (acceso de datos online).

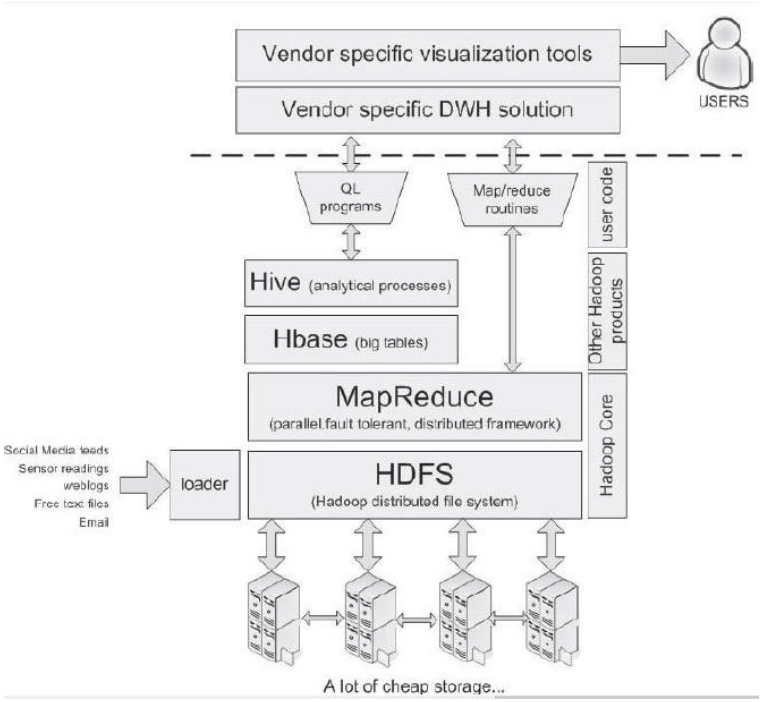


Figura N° 7: Arquitectura de Hadoop

1. HDFS: Hadoop Distributed File System, es un sistema de ficheros basado en la distribución de la información en distintas máquinas que pueden ser geográficamente muy distantes, conectadas entre sí mediante una red de modo transparente al usuario.
2. Map Reduce: Es un paradigma de programación dividido en dos fases: Map y Reduce. tiene la capacidad de dividir una petición por parte de un cliente en otros muchas partes y encargar el trabajo a múltiples nodos que funcionan en paralelo. MapReduce es el principal encargado de la gestión de recursos y procesamiento de datos, su arquitectura general es de la forma:
   - ✓ Map: Su función consiste básicamente en el mapeo de la información entrante. Esta fase tiene como entrada la información y como salida un par [clave: valor] que será la entrada de la siguiente fase.
   - ✓ Reduce: Esta fase es la encargada de realizar el procesamiento de la información recibida, ya mapeada en el paso anterior. Tiene como entrada el par [clave: valor] obtenido de la fase anterior y como salida otro par [clave: valor].

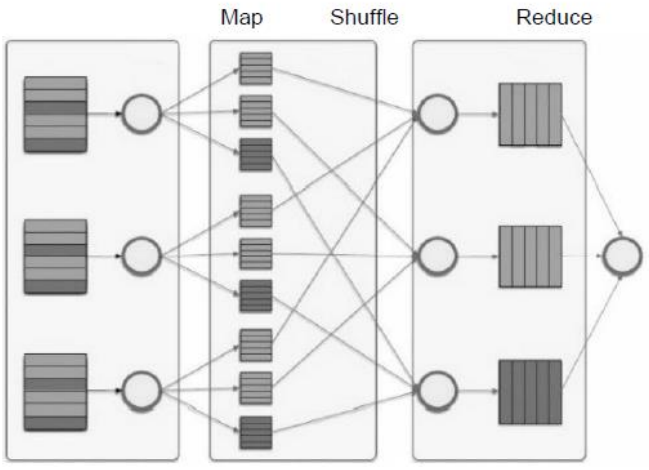


Figura N° 8: Arquitectura de Funcionamiento de MapReduce

En la figura un programa de solución o trabajo en Hadoop ejecuta 4 pasos principales:

i) División de datos donde múltiples fracciones de datos son entregadas a cada uno de los mapeadores,

ii) Map: Donde se ejecutan funciones para procesar los datos, con la identificación de elementos relevantes y enviarlos a la etapa de organización.

iii) Organización de los datos y agrupar resultados intermedios, entonces distribuirlos hacia la etapa de reducción, y

iv) Reduce: Donde se ejecutan las funciones para compactar y resumir los resultados a ser escritos en disco. En la práctica, objetos mapeadores y reductores ejecutan funciones map y reduce, respectivamente. Así, objetos mapeadores son responsables de procesar pares clave-valor para producir un conjunto de pares clave-valor intermedios; y luego, objetos reductores procesar y resumir el conjunto de valores intermedios junto con su clave compartida.

Igualmente, Vidal, Bustamante, Lapo y Nuñez (2018) resumen que como tal, MapReduce es un enfoque de computación para trabajar con grandes volúmenes de datos (Big Data) en un entorno distribuido, con altos niveles de abstracción y con el uso ordenado de funciones Map y Reduce, la primera de ellas para el mapeo o identificación de datos relevantes y la segunda para resumir datos y resultados finales. De la misma manera los autores indican que MapReduce, es una metodología de programación dada a conocer por Google, para la computación distribuida sobre grandes cantidades de datos.

3. Hive; Es una tecnología desarrollada en Facebook que convierte Hadoop en un almacén de datos completo con un dialecto de SQL para realizar consultas, HiveQL es un lenguaje declarativo. Hive se da cuenta de cómo construir un flujo de datos para

lograr que resultado se espera, en Hive se requiere un esquema, pero no se limitan a uno solo.

3.7.2. Rhaddop

Es la combinación de dos tecnologías como es R y Hadoop, las cuales se complementan de una forma ventajosa, porque permite el análisis y la visualización de grandes volúmenes de datos. RHadoop es definido como una colección de paquetes R que permiten que el usuario pueda procesar grandes volúmenes de datos en Hadoop. RHadoop es considerado también como "un puente entre R, un lenguaje y entorno para explotar estadísticamente conjunto de datos, y Hadoop, un marco que permite el procesamiento distribuido de grandes conjuntos de datos a través de clústers de computadores".

3.8. Data Scientist

Marco Bressan, Chairman, CEO del BBVA tiene como el objetivo principal de su departamento el capturar valor de los datos (Twitter) que tiene el banco. Con una cantidad de datos importante de los clientes, de sus obligaciones, buscando que en primer lugar, centralizar esos datos, poder mirarlos y poder extraer conocimiento, extraer información de esos datos y traducir ese conocimiento e información en nuevos servicios, en mejores servicios para nuestros clientes, etc. Por ello la aplicación más importante de Big Data (Facebook) es extraer información de nuestros clientes a partir de lo que hacen, ya no es hacer una encuesta, salir y preguntarle al cliente que le gustaría, sino realmente ver, por ejemplo, cómo utiliza su cuenta, qué tipo de movimientos hace, qué tipo de vida tiene, y tratar de darle a ese cliente un mejor servicio.

En esencia lo que persigue el BBVA según el CEO: "El objetivo es transformar información en servicios para clientes".

A continuación el autor hace una evaluación y análisis sobre el rol del data scientist en su día a día. El día a día de un equipo que trabaja en data análisis se puede definir como un esfuerzo multidisciplinar, porque es muy difícil que una persona aislada delante de un ordenador pueda obtener toda esa información. La clave del equipo está en las personas que trabajan juntas en varias disciplinas.

Como tal no solo se requieren conocimientos tecnológicos, estadísticos o matemáticos, también científicos de diferentes áreas: ya sean sociales, geográficas, geopolíticas, económicas, por eso, aunque pueda parecer que estamos trabajando con un ordenador, se trata de matemática y mucha estadística. Es muy divertido porque es un trabajo que, sobre todo, explota la curiosidad, creatividad, exploración, prueba y error. Hay que trabajar con datos, meterse dentro de ellos y tratar de encontrar patrones e información que pueda llegar a ser relevante.

Figura N° 9: Claves del Big Data

Figura N° 10: Metodología y Consejos

CAPITULO IV: ANALISIS E INTERPRETACION DE DATOS

4.1. Análisis de los datos

Con la finalidad de poder cumplir con los objetivos específicos de la investigación como lo son: a) Descubrir el conocimiento que se tiene en los estudiantes de la FIS-UNICA sobre el BIGDATA, b) Motivar al conocimiento del BIGDATA en los estudiantes de la FIS-UNICA y c) Profundizar en los conocimientos sobre el BIGDATA en la FIS-UNICA.

Para ello se diseñó un formulario virtual para conocer dicho conocimiento, motivar al conocimiento y profundizar con la presente investigación todo lo referente al Big Data (Anexo 02). Cabe resaltar que el cuestionario fue aplicado a los estudiantes de la FIS-UNICA del octavo y decimo ciclo, que ya se encuentran con conocimientos avanzados por estar próximos a acabar su formación profesional y que aproximadamente son los resultados que se detallan después de la realización del cuestionario.

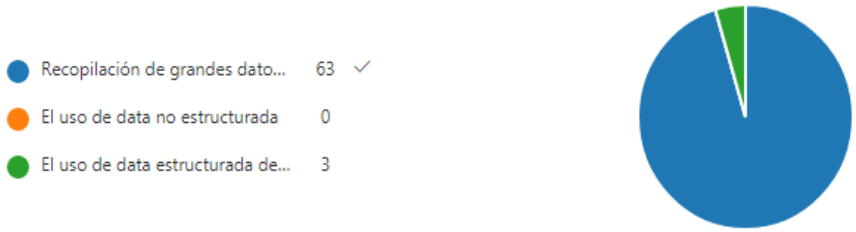
Figura N° 11: Resultados de la encuesta de opinión

La figura sobre las respuestas indica que han sido 69 estudiantes los que han accedido a responder el cuestionario.

1. ¿Qué concepto define con mayor precisión a BIG DATA?

Un 95 % de los usuarios que completaron el cuestionario (63 de 66) respondió correctamente a esta pregunta.

Más detalles



Interpretación: los resultados de la interrogante nos dan de los 69 (95%) estudiantes que han respondido, 63 de ellos han acertado en su respuesta, se destaca que 3 estudiantes no han respondido a esa pregunta y 3 han marcado incorrecto la respuesta.

2. ¿Cual es el objetivo del Big Data?

Un 55 % de los usuarios que completaron el cuestionario (35 de 64) respondió correctamente a esta pregunta.

Más detalles

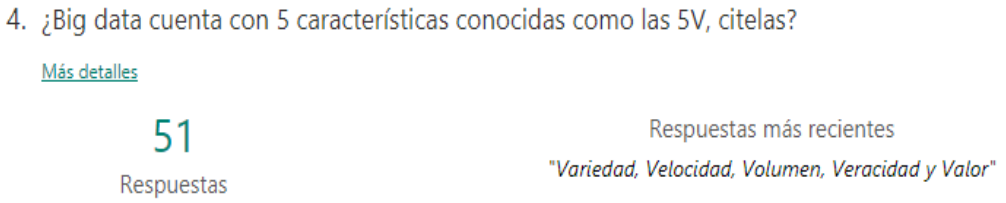
Interpretación: al respecto sobre el objetivo del Big Data, de los 69 estudiantes 35 (55%) han acertado con su respuesta, 29 no han respondido correctamente la pregunta y solo 5 estudiantes no ha respondido a la pregunta.

3. ¿Cuál son los beneficios del Big Data?

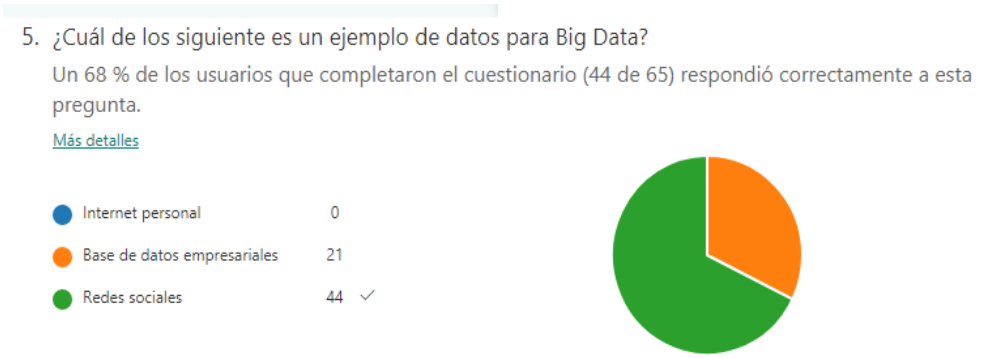
Un 47 % de los usuarios que completaron el cuestionario (30 de 64) respondió correctamente a esta pregunta.

Más detalles

Interpretación: en relación a los beneficios del Big Data, de los 69 estudiantes que han ingresado al cuestionario, 30 (47%) de ellos han acertado con la respuesta, 34 de ellos no han acertado con la respuesta y 5 de ellos no han respondido a la pregunta.



Interpretación: en relación sobre las características del Big Data, de los 69 estudiantes, 51 (100%) han respondido favorablemente a la pregunta, y 18 estudiantes no han respondido.

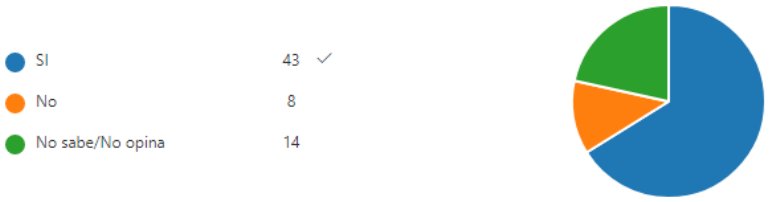


Interpretación: para la pregunta sobre ejemplos de Big Data, de los 69 que han ingresado al cuestionario, 44 (68%) de ellos han acertado con la respuesta, mientras que 21 no ha respondido correctamente y 4 estudiantes no han respondido.

6. ¿Los datos no estructurados, utilizan bases de datos NoSQL?

Un 66 % de los usuarios que completaron el cuestionario (43 de 65) respondió correctamente a esta pregunta.

Más detalles

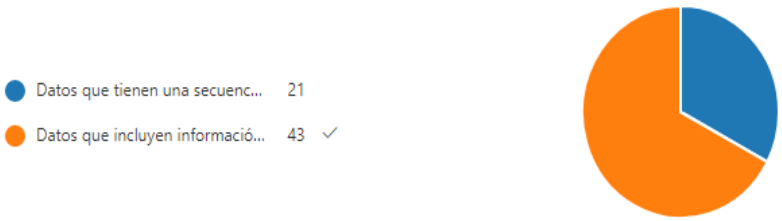


Interpretación: a la interrogante sobre datos estructurados, de los 69 estudiantes que han ingresado al cuestionario, 43 (66%) de ellos han acertado con la pregunta, mientras que 8 no y 14 que no sabe/no opina, e igual hay 4 estudiantes que no han respondido.

7. Se consideran datos no estructurados

Un 67 % de los usuarios que completaron el cuestionario (43 de 64) respondió correctamente a esta pregunta.

Más detalles



Interpretación: A la pregunta sobre los datos no estructurados, de los 69 que han ingresado al cuestionario, 43 (67%) de ellos han respondido favorablemente, mientras que 21 no a respondido correctamente, mientras que 15 no han respondido a la pregunta.

8. ¿Cuál es el flujo de trabajo relacionado al Big Data?

Un 42 % de los usuarios que completaron el cuestionario (27 de 64) respondió correctamente a esta pregunta.

Más detalles

Interpretación: en relación al flujo de trabajo del Big Data, de los 69 estudiantes que han ingresado al cuestionario, 27 (42%) han acertado con la respuesta, mientras que 37 de ellos no, destacando que existen 5 estudiantes que no han respondido a ella.

9. Cite dos aplicaciones que conozca que se utilicen para BIg data

Más detalles

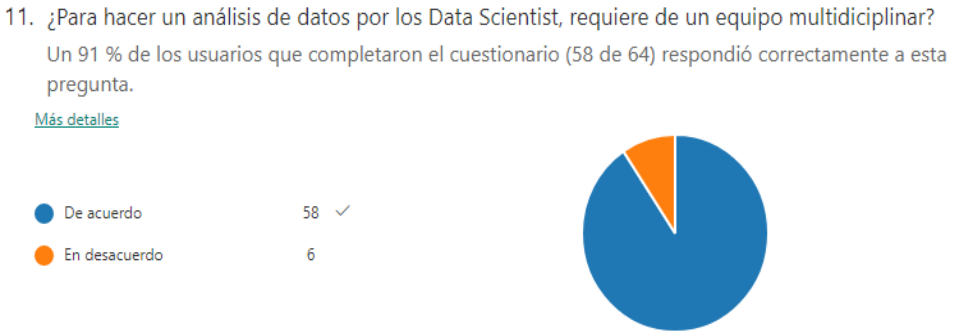
Interpretación: en la pregunta de los 49 (100%) estudiantes que han respondido a la pregunta sobre alguna aplicación que conozca, solo 2 respuestas satisfacen con la pregunta. 47 estudiantes han dado respuestas sobre fuentes de datos para Big Data como Facebook, twuiter, Instagram, etc. 20 de ellos no han respondido.

10. ¿En que consiste un Data Science, relacionado con el Big data?

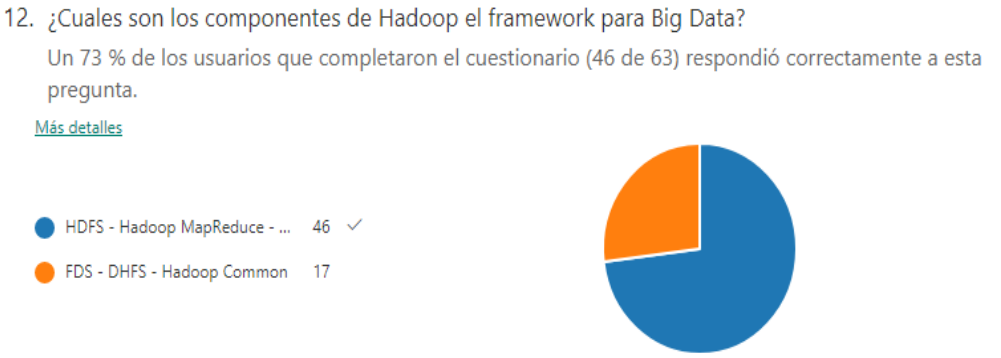
Un 70 % de los usuarios que completaron el cuestionario (44 de 63) respondió correctamente a esta pregunta.

Más detalles

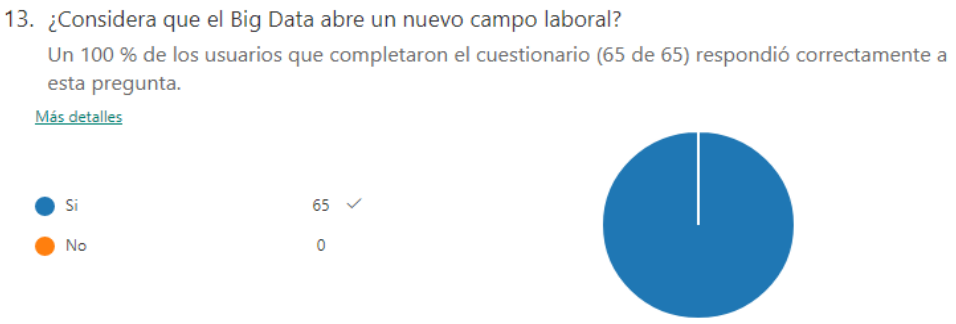
Interpretación: en relación al Data Science 44 (70%) de los 69 que han ingresado al cuestionario han respondido favorablemente, mientras que 19 de ellos no, destacando que 6 de ellos no han respondido a la pregunta.



Interpretación: en relación a la multidisciplina de los Data Scientits, 58 (91%) de ellos responden favorablemente, mientras que 6 de los estudiantes no acierta con la respuesta y 5 de ellos no han respondido a la pregunta.



Interpretación: en relación a la pregunta sobre Hadoop, 46 (73%) de ellos respondió favorable, mientras que 17 de los estudiantes no acertó con la respuesta, mientras que 3 de los estudiantes no respondieron a la pregunta.



Interpretación: en relación que el Big Data abre un nuevo campo laboral, 65 (100%) de los estudiantes, afirma que, si efectivamente abre un nuevo campo laboral, 5 de los estudiantes no han respondido a la pregunta.

A continuación, en la tabla siguiente presentamos un resumen de los resultados de las preguntas del cuestionario.

Tabla N° 3
Resumen de resultado del cuestionario

N° Preg	Preguntas respondidas	Preguntas acertadas	% de aciertos	Preguntas no respondidas
1.	66	63	95	3
2.	64	35	55	5
3.	64	30	47	5
4.	51	51	100	18
5.	65	44	68	4
6.	65	43	66	4
7.	64	43	67	5
8.	64	27	42	5
9.	49	2	4	20
10.	63	44	70	6
11.	64	58	91	5
12.	63	46	73	6
13.	51	51	100	18
Media	61	41	68	8

Interpretación: los resultados obtenidos del cuestionario en la cual de los 69 estudiantes que han ingresado al cuestionario, se tiene que en promedio 61 (88.41%) de ellos han respondido a las preguntas, en donde de estos 61 que ha respondido han acertado en sus repuestas 41 de ellos y que en promedio representan un 68% de aciertos. Tenerse en cuenta que en promedio 8 de los estudiantes no han respondido a las preguntas.

En la tabla se debe tener en cuenta que los porcentajes que se han calculado por el sistema del Form del Office 365, se establecen en relación a las preguntas respondidas y no se toma en cuenta las preguntas no respondidas.

CAPITULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

Culminado nuestro estudio, el mismo que no solo ha ampliado mis conocimientos sobre el Big Data, como una tendencia en el mundo, el empleo efectivo de los datos que se están generando en el mundo por diversas organizaciones y cumplido el objetivo establecido, llegamos a las siguientes conclusiones:

1. Existe un buen conocimiento en general sobre el Big Data en los estudiantes de la FIS-UNICA, que se reflejan en el 68% en promedio de aciertos de los estudiantes que han respondido al cuestionario.
2. Bajo esta evaluación solo la interrogante sobre aplicaciones para Big Data los resultados no han sido favorables ya que solo 4% de los estudiantes que han respondido han acertado a la pregunta, las respuestas demuestran que hay confusión entre las aplicaciones como Hadoop, MapReduce, y se confunde con las fuentes de datos para Big Data como Facebook, Twitter, Instagram u otras redes y fuentes de datos.
3. El punto anterior nos lleva a concluir que si bien es cierto, los estudiantes tienen conocimiento general sobre el Big Data, este es reducido sobre las herramientas y aplicaciones que se utilizan para el manejo de grandes cantidades de datos.
4. Es importante resaltar que al haber un conocimiento promedio del 68% favorable sobre Big Data, es de destacar que el 100% de los estudiantes que han respondido están conscientes que esta tendencia del aprovechamiento de la gran cantidad de datos que se están produciendo en el mundo y en las organizaciones abren un nuevo camino para aprovechar profesionalmente.

5.2. Recomendaciones

Como consecuencia de la culminación del estudio, es importante las recomendaciones que se derivan de ella y que a continuación, serán puesta a disposición de la comunidad para su aprovechamiento.

1. Los resultados de la investigación, orientada a conocer sobre el conocimiento general sobre el Big Data nos han permitido identificar que el mundo se está orientando al aprovechamiento de los datos que se están generando, sean estos estructurados o no estructurados, es por ello que se recomienda que se aproveche dicha orientación para que en la FIS-UNICA, se incorpore en el plan de estudios de especialidad Big Data, tal es el caso de otras universidades.
2. Por otro lado el estudio no profundiza en el conocimiento sobre el uso y el aprovechamiento de esta nueva tendencia, por lo que se recomienda hacer nuevos estudios en profundidad sobre las aplicaciones como Haddop, MapReduce y metodologías aplicables a Big Data.
3. A nivel de la Universidad Nacional San Luis Gonzaga, ésta tiene más de 60 años de existencia, en la cual se han generado grandes cantidades de datos e información que están allí esperando a poder ser aprovechadas, y sin embargo el uso de las tecnologías a la fecha viene siendo incipiente; por lo que se recomienda hacer un estudio del aprovechamiento del Big Data, cuyos resultados sean aprovechados por la Universidad para mejorar sus servicios hacia los estudiantes.
4. Finalmente siendo la FIS-UNICA la facultad que debe liderar el cambio en la universidad, y con los resultados del estudio en la que los estudiantes tienen un conocimiento aceptable en relación al Big Data, se recomienda que por medio de los estudiantes de los últimos ciclos se proponga hacer un piloto sobre Big Data, orientado al aprovechamiento de los datos académicos.

REFERENCIA BIBLIOGRAFICAS

Castaño Mar y Ortega Felipe (20 Octubre 2016). Técnicas de Análisis de Datos y Bigdata. Disponible en: https://www.youtube.com/watch?time_continue=115&v=Ur8I7UwMu6A

DELL EMC (s.f.). ¿Qué es Big data?. Disponible en: https://www.dellemc.com/es-pe/big-data/definitions.htm.

GUERRERO LÓPEZ, FABIÁN ANDRÉS y RODRÍGUEZ PINILLA, JORGE EDUARDO (2013). DISEÑO Y DESARROLLO DE UNA GUÍA PARA LA IMPLEMENTACIÓN DE UN AMBIENTE BIG DATA EN LA UNIVERSIDAD CATÓLICA DE COLOMBIA. Disponible en: https://repository.ucatolica.edu.co/.../DISEÑO%20Y%20DESARROLLO%20DE%20U

Hernández, Fernández y Baptista (2010). Metodología de la Investigación. 5ta Ed. México. Ed. Mc Graw Hill

Hernández Leal, Emilcy Juliana (2016). Aplicación de técnicas de análisis de datos y administración de Big Data ambientales. Disponible en: http://bdigital.unal.edu.co/54512/1/1090175695.2016.pdf

IBM (s.f.). **Mucho más que un simple cloud: los servidores de IBM Cloud se ejecutan en cualquier nivel. Disponible en:** . https://www.ibm.com/cloud-computing/bluemix/es/info/fast-cloud-servers?S_PKG=&cm_mmc=Search_Google-_-Cloud_Cloud+Platform-_-EP_PE-_-cloud+computing_Exact_&cm_mmca1=000016GC&cm_mmca2=10004026&cm_mmca7=9073221&cm_mmca8=aud-311016886972:kwd-

295170339346&cm_mmca9=7122dcd3-b54e-494e-9860-7c1b787397f4&cm_mmca10=267200738138&cm_mmca11=e&mkwid=7122dcd3-b54e-494e-9860-7c1b787397f4|636|293121&cvosrc=ppc.google.cloud%20computing&cvo_campaign=000016GC&cvo_crid=267200738138&Matchtype=e

La Nacion (19 julio 2011). ¿Qué es la nube, para que sirve y cuales son los servicios que ten´ss que conocer?. Disponible en: https://www.lanacion.com.ar/1389864-que-es-la-nube-para-que-sirve-y-cuales-son-los-servicios-que-tenes-que-conocer

Laverde Salazar, María Fernanda (2015). Diseño de un curso teórico y práctico sobre: Big Data. Disponible en: http://repositorio.uchile.cl/handle/2250/139432

López García, David (2013). Análisis de las posibilidades de uso de Big Data en las organizaciones. Disponible en: https://repositorio.unican.es/xmlui/bitstream/handle/10902/4528/TFM%20-%20David%20L%C3%B3pez%20Garc%C3%ADaS.pdf?sequence=1

…

LOSTAUNAU FUENTES, MIGUEL (2015). PROBLEMAS DE USO DE DATOS SOBRE EL CRIMEN EN LOS INFORMES DEL ESTADO. Disponible en: http://tesis.pucp.edu.pe/repositorio/bitstream/handle/123456789/6696/LOSTAUNAU_FUENTES_MIGUEL_PROBLEMAS.pdf?sequence=1

Manso, Fernando (2015). Análisis de Modelos de Negocios Basados en BIG DATA para Dispositivos Móviles. Disponible en: http://repositorio.udesa.edu.ar/jspui/bitstream/10908/10920/1/%5BP%5D%5BW%5D%20T.%20M.%20Ges.%20Manso%2C%20Fernando.pdf

Ortega Felipe, Gómez Victoria, Baquero José Carlos & Bengamin Richard (20 Octubre 2016). Tendencias futuras en análisis de Bigdata. Disponible en: https://www.youtube.com/watch?time_continue=14&v=AIUTbwEz5TI

Ortega Felipe (9 Noviembre 2016). Data Mining, tendencias en análisis y visualización de datos. Disponible en: https://www.youtube.com/watch?time_continue=46&v=_IyQ3ocYbDo

Rojas García, José Antonio (2016). Propuesta de un modelo de negocio basado en big data que facilite la integración de los datos de las personas naturales y de soporte a las políticas de e-government en el Perú, apoyado en una empresa de logística integral. Disponible en: https://repositorioacademico.upc.edu.pe/handle/10757/620937.

SAS, The Power to Know (s.f.). La Minería de Datos de la A a la Z. disponible en: https://www.sas.com/es_cl/campaigns/analytics/mineria-datos.html?gclid=EAIaIQobChMItKXapqih3QIVRSaGCh3AcAsvEAAYASAAEgK3LPD_BwE

Webempresa2.0 (s.f.). Las 30 Redes sociales más Utilizadas. Disponible en: https://www.webempresa20.com/blog/las-30-redes-sociales-mas-utilizadas.html

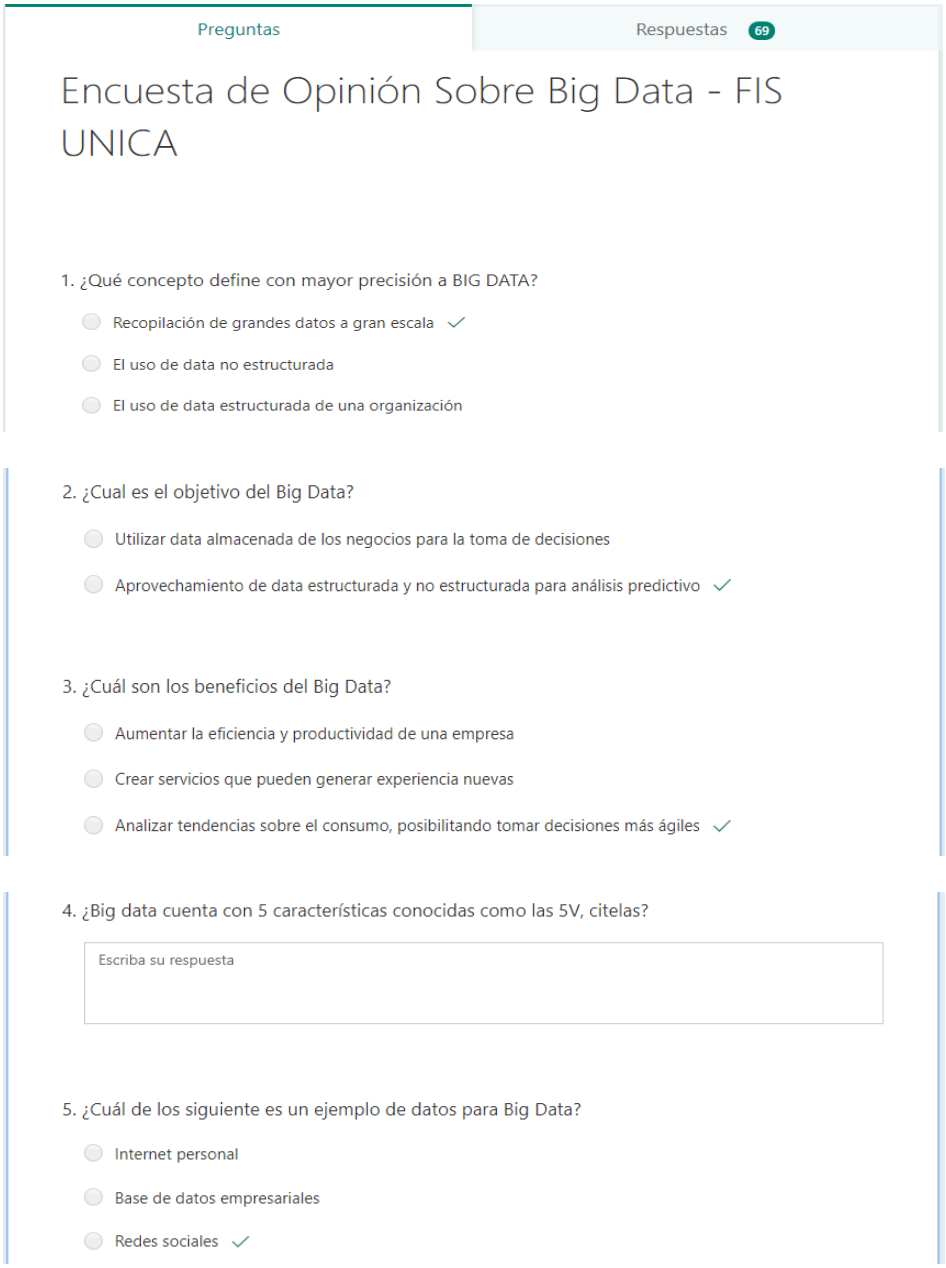
Wikipedia (18 octubre 2017). Redes Sociales en Internet, disponible en: https://es.wikipedia.org/wiki/Redes_sociales_en_Internet

Anexo 01: MATRIZ DE CONSISTENCIA

PROBLEMA	OBJETIVO	METODOLOGIA	TECNICAS	HERRAMIENTAS
¿Cuál es el grado de conocimiento del BIGDATA, y que herramientas están asociadas a su aplicación, en los estudiantes de la FIS-UNICA?	entender con mayor amplitud al BIGDATA, como nueva tecnología para grandes volúmenes de datos, que herramientas están relacionadas con él, así como el alcance de los datos para sea aprovechado por el BIGDATA. OBJETIVOS ESPECIFICOS: a) descubrir el conocimiento que se tiene en los estudiantes de la	Tipo de investigación Transversal Nivel de Investigación Exploratoria Diseño no Experimental Población: los estudiantes de la Facultad de Ingeniería de Sistemas Muestra: Estudiantes de las asignaturas de Teoría y Diseño de Base de Datos e Implementación de Base de Datos (2018-I).	Entrevista Encuesta Análisis documental	Guía de Entrevista Cuestionario Fichas documentales

	FIS-UNICA sobre el BIGDATA, b) motivar al conocimiento del BIGDATA en los estudiantes de la FIS-UNICA, c) Profundizar en los conocimientos sobre el BIGDATA en la FIS-UNICA.			

Anexo 02: Cuestionario Virtual sobre conocimiento sobre Big Data



Preguntas | Respuestas 69

Encuesta de Opinión Sobre Big Data - FIS UNICA

1. ¿Qué concepto define con mayor precisión a BIG DATA?
   - Recopilación de grandes datos a gran escala ✓
   - El uso de data no estructurada
   - El uso de data estructurada de una organización

2. ¿Cual es el objetivo del Big Data?
   - Utilizar data almacenada de los negocios para la toma de decisiones
   - Aprovechamiento de data estructurada y no estructurada para análisis predictivo ✓

3. ¿Cuál son los beneficios del Big Data?
   - Aumentar la eficiencia y productividad de una empresa
   - Crear servicios que pueden generar experiencia nuevas
   - Analizar tendencias sobre el consumo, posibilitando tomar decisiones más ágiles ✓

4. ¿Big data cuenta con 5 características conocidas como las 5V, citelas?

   Escriba su respuesta

5. ¿Cuál de los siguiente es un ejemplo de datos para Big Data?
   - Internet personal
   - Base de datos empresariales
   - Redes sociales ✓

6. ¿Los datos no estructurados, utilizan bases de datos NoSQL?

- SI ✓
- No
- No sabe/No opina

7. Se consideran datos no estructurados

- Datos que tienen una secuencia lógica y se pueden almacenar
- Datos que incluyen información sobre fotos, videos, SMS, Twiter, Facebook ✓

8. ¿Cuál es el flujo de trabajo relacionado al Big Data?

- Extrapolación - Entendimiento - Reproducción ✓
- Modelo - Precisión - Big data

9. Cite dos aplicaciones que conozca que se utilicen para BIg data

Escriba su respuesta

10. ¿En que consiste un Data Science, relacionado con el Big data?

- Conocimiento por medio de los datos
- Ciencia de los datos
- Estudio sistemático de la información

10. ¿En que consiste un Data Science, relacionado con el Big data?

- Conocimiento por medio de los datos
- Ciencia de los datos
- Estudio sistemático de la información
- Recolección, preparación, análisis, visualización, administración y preservación de grandes volúmenes de información ✓

11. ¿Para hacer un análisis de datos por los Data Scientist, requiere de un equipo multidiciplinar?

- De acuerdo ✓
- En desacuerdo

11. ¿Para hacer un análisis de datos por los Data Scientist, requiere de un equipo multidiciplinar?

- De acuerdo ✓
- En desacuerdo

12. ¿Cuales son los componentes de Hadoop el framework para Big Data?

- HDFS - Hadoop MapReduce - Hadoop Common ✓
- FDS - DHFS - Hadoop Common

13. ¿Considera que el Big Data abre un nuevo campo laboral?

- Si ✓
- No