EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Estudio de mapeo sistemático en el problema de la variedad en sistemas Big Data

presentada por

Ing. Daniel Ramírez Gervacio

como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Director de tesis

Dr. Juan Carlos Rojas Pérez

Codirector de tesis

Dra. Olivia Graciela Fragoso Díaz

Cuernavaca, Morelos, México.Enero del 2021.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 08/diciembre/2020

OFICIO No. DCC/121/2020
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

C. DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del **C. Ing. Daniel Ramírez Gervacio**, con número de control M18CE066, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Estudio de Mapeo Sistemático en el problema de Variedad en Sistemas Big Data"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

Dr. Juan Carlos Rojas Pérez
Doctor en Ciencias en Ciencias de la Computación
6099372
Director de tesis

Dra. Olivia Graciela Fragoso Díaz
Doctora en Ciencias en Ciencias de la Computación
7420199
Co-directora de tesis

Dr. René Santaolaya Salgado
Doctor en Ciencias de la Computación
4454821
Revisor 1

M.C. Mario Guillén Rodríguez
Maestro en Ciencias con Especialidad en Sistemas Computacionales
7573768
Revisor 2

C.c.p. Depto. Servicios Escolares
Expediente / Estudiante
JGGS/lmz

Cuernavaca, Morelos 14/diciembre/2020
OFICIO No. SAC/ 285/2020
Asunto: Autorización de impresión de tesis

DANIEL RAMÍREZ GERVACIO
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado ***"Estudio de Mapeo Sistemático en el problema de Variedad en Sistemas Big Data",*** ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO SUBDIRECCIÓN ACADÉMICA

C.c.p. M.E. Guadalupe Garrido Rivera. Jefa del Departamento de Servicios Escolares
Expediente
CMAZ/CHG

Dedicatoria

Dedico este trabajo principalmente a Dios, por haberme dado la vida y permitirme el haber llegado hasta este momento tan importante de mi formación profesional.

A mis padres, por su amor, trabajo y sacrificio en todos estos años, gracias por haberme forjado como la persona que soy en la actualidad; muchos de mis logros se los debo a ustedes.

A mi hermano Héctor, por siempre estar presente en cada momento, por todo el apoyo y cariño a lo largo de mi vida.

A mis abuelos por haberme enseñado muchas cosas vitales para la vida y que hora son ángeles en mi vida. Sé que se encuentran muy orgullosos de su nieto.

A mí, por la preseverancia, constancia, esfuerzo y dedicación para lograr este trabajo.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), por proporcionarme el apoyo económico por medio de una beca durante el transcurso de la maestría, la cual me ayudó a culminar esta meta.

Al Tecnológico Nacional de México (TecNM) / Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), por darme la oportunidad de realizar mis estudios de maestría y por ser una institución académica de excelencia.

Especialmente a mi director de tesis, Dr. Juan Carlos Rojas Pérez, por su orientación, apoyo y enseñanzas para llevar a cabo este trabajo de tesis, sobretodo su paciencia y tiempo. ¡Muchas gracias!

A mi Codirectora de tesis, Dra. Olivia G. Fragoso Díaz, por haberme brindado su apoyo, tiempo, paciencia y por su orientación en el desarrollo de mi trabajo de tesis.

A mis revisores: Dr. René Santaoyala Salgado y M.C. Mario Guillen Rodríguez, por sus observaciones y sugerencias a lo largo de la maestría para realizar un trabajo de calidad.

A mis padres por haberme dado la oportunidad de superarme en la vida. Siempre han sido un ejemplo para mí. Gracias a ello estoy alcanzando mis metas con mucho orgullo. Les debo un eterno agradecimiento y mi retribución total por su gran amor.

A mi Hermano y su esposa, por haberme motivado a seguir adelante y por ayudarme cuando más lo necesite.

A mi novia Ana Laura, que me ha enseñado a ser una mejor persona, por su apoyo, cariño, amistad y por siempre estar conmigo en los mejores y peores momentos.

A mi amigo Juan Carlos Carbajal Martínez, por haberme apoyado cuando más lo necesite, por tu amistad y por todas las enseñanzas que me has brindado.

A mis amigos y compañeros que conocí durante mi estancia en el CENIDET: Ana, Nelida, Angélica, Jahaziel, Moisés, Irving, Diego, Sergio, Araceli, Edgar, José, Carlos, Axel, Viterbo, Capistran, Orlando, Ricardo y Marisol.

Resumen

Actualmente el término Big Data no solo hace referencia a la gran cantidad de datos, sino que también hace referencia a la información que no se puede procesar o analizar utilizando procesos o herramientas tradicionales debido a 3 características: el volumen, la velocidad y la variedad. El constante aumento de información que se genera en la actualidad y el desarrollo de nuevas tecnologías han traído consigo un crecimiento exponencial de nuevos tipos de datos, los cuales son difíciles de procesar con herramientas o frameworks previos al surgimiento de Big Data. Sin embargo, el mayor problema en el análisis de Big Data no es la gran cantidad de datos (Volumen), sino la dificultad de analizar los diferentes tipos de datos (Variedad). El presente documento tiene como objetivo evidenciar el estatus en cuanto al uso de herramientas, frameworks y metodologías utilizadas en el problema de la variedad en Big Data. Para esto se llevó a cabo un estudio de mapeo sistemático, que inició con una pregunta de investigación principal hasta llegar a la búsqueda, selección y clasificación de estudios relevantes.

Una de las finalidades de este trabajo de investigación es permitir guiar futuras investigaciones, de tal forma que pueda ser utilizado como referencia de línea base.

Abstract

Today the term Big Data not only refers to the large amount of data, but also to information that cannot be processed or analyzed using traditional processes or tools due to 3 characteristics: volume, velocity and variety. The constant increase of information that is generated nowadays and the development of new technologies have brought an exponential growth of new types of data, which are difficult to process with tools or frameworks previous to the emergence of Big Data. However, the biggest problem in the analysis of Big Data is not the large amount of data (Volume), but the difficulty of analyzing different types of data (Variety). The present document aims to evidence the currently status in the use of tools, frameworks and methodologies used in the problem of variety in Big Data. For this purpose, a systematic mapping study was carried out, starting with a main research question and ending with the search, selection and classification of relevant studies.

One of the purposes of this research work is to guide future research, so that it can be used as a baseline.

Índice General

Índice de Figuras

Índice de Tablas

Índice de Gráficas

Acrónimos

3v´s:	Características principales de Big Data las cuales son el Volumen, la Velocidad y la Variedad.
AAMBDA:	An Architecture and Methods for Big Data Analysis. (Artículo)
AIMBDMAV:	An Iterative Methodology for Big Data Management, Analysis and Visualization. (Artículo)
API:	Application Programming Interface (Interfaz de programación de aplicaciones).
ASMHBDV:	A Storage Model for Handling Big Data Variety. (Artículo)
BDMPT:	Big Data: Methods, Prospects, Techniques. (Artículo)
BDPMP(IMAGS):	Big data preprocessing: methods and prospects. (Artículo)
BHBDCMA:	Beyond the hype: Big Data concepts, methods, and analytics. (Artículo)
CCTV:	Closed circuit television (circuito cerrado de televisión).
Cenidet:	Centro Nacional de Investigación y Desarrollo Tecnológico.
CEP:	Complex Event Processing (Procesamiento de eventos complejos).
CONRICYT:	Consorcio Nacional de Recursos de Información Científica y Tecnológica.
CQL:	Continuous Query Language (Lenguaje de consultas continuo).
CRF	Conditional Random Field (Campo aleatorio condicional).
CSS:	Cascading Style Sheets (Hojas de Estilo en Cascada).
CSV:	Comma Separated Values (Valores Separados por Comas).
DB:	Data Base (Base de datos).
EST:	Datos Estructurados.
ETL:	Extract, Transform and Load (Extraer, transformar y cargar).
FC-BESSSED:	Framework for Concept-Based Exploration of Semi-Structured. (Artículo)

FERIUUBD:	Framework for Extracting Reliable Information from Unstructured Uncertain Big Data. (Artículo)
FHDHC:	Framework to Handle Data Heterogeneity Contextual. (Artículo)
FIBD:	Framework of Integrated Big Data. (Artículo)
FUDA:	Framework for Unstructured Data Analysis. (Artículo)
FWK	Framework
HDFS:	Hadoop Distributed File System (Sistema de ficheros distribuido de Hadoop).
HTA	Herramienta.
HTML:	HyperText Markup Language (Lenguaje de Marcado de Hipertextos).
IBM:	International Business Machines (Empresa).
ID:	Identificador.
IDE´s:	Integrated Development Environment (Entorno de desarrollo integrado).
IoBD	Internet of Big Data (Internet del Big Data).
IoT:	Internet of Things (Internet de Las Cosas).
JSON:	JavaScript Object Notation (Notación de objetos de JavaScript).
KDD	Knowledge Discovery in Databases (Descubrimiento de conocimiento en bases de datos).
MDANJHVD:	Multi-model Databases: A New Journey to Handle the Variety of Data. (Artículo)
MTG	Motodología
N/A:	No Aplica.
N-EST:	Datos No estructurados.
RDBMS:	Relational Database Management System (Sistema de gestión de bases de datos relacionales).
RDD:	Resilient Distributed Dataset (Conjunto de datos distribuidos resilientes).

RQ:	Research Question (Pregunta de investigación).
RTF:	Rich Text Format (Formato de texto enriquecido).
S-EST:	Datos Semiestructurados.
SIECRF	Structure Information Extraction model based on CRF (Modelo de extracción de información de estructura basado en CRF).
SLR	Systematic Literature Review (Revisión sistemática de la literatura)
SMS:	Systematic Mapping Study (Estudio de Mapeo Sistemático).
SQL:	Structured Query Language (Lenguaje de consulta estructurada).
TSV:	Tab Separated Values (Valores separados por tabulaciones).
TXT:	Textfile (Archivo de texto).
VAR	Variedad.
VEL	Velocidad.
VOL	Volumen.
XLS:	eXceL Spreadsheets (Hojas de cálculo Excel).
XML:	Extensible Markup Language (Lenguaje de Marcado Extensible).
ZIP:	Zone Information Protocol (Protocolo de Información de Zona).

Capítulo 1

Introducción

El constante avance de las tecnologías de información, ha permitido un crecimiento avanzado en la cantidad de datos generados desde diferentes fuentes, tales como, redes sociales, dispositivos móviles, sensores, sondas espaciales, sistemas de predicción del clima, sistemas de geo-posicionamiento, entre otros, caracterizados por tratarse de datos, en su mayoría, sin estructura. (K. Stephen, 2013)

Actualmente Big Data se está convirtiendo en el próximo recurso natural que explotar; y esto representa por un lado un gran reto, pero también una oportunidad para las organizaciones que sepan sacar provecho de estos datos. Para sumarse a esta oportunidad, las ciudades deben adoptar soluciones que proporcionen capacidades analíticas para convertirlas en conocimiento y así mejorar la gestión urbana y la toma de decisiones. (Felipe, 2015)

Cuando se habla de Big Data se hace referencia a grandes conjuntos o combinaciones de datos cuyo tamaño (volumen), complejidad (variedad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías, herramientas y frameworks convencionales.

La variedad se refiere a la diversidad que existe en la estructura en un conjunto de datos. Los avances tecnológicos permiten a las empresas utilizar diversos tipos de datos estructurados, semiestructurados y no estructurados. Los datos estructurados, que componen solo el 5% de todos los datos existentes, se refieren a los datos que se encuentran en las hojas de cálculo o bases de datos relacionales (Cukier, 2010). El texto, imágenes, audio y video son ejemplos de datos no estructurados, que a veces requieren una organización estructural para el análisis.

Hoy en día las organizaciones han estado acumulando datos no estructurados de fuentes internas (por ejemplo, datos de sensores) y fuentes externas (por ejemplo, redes sociales). Sin embargo, la aparición de nuevas tecnologías de gestión de datos y análisis, permiten a las organizaciones aprovechar los datos en sus procesos de negocio, es el aspecto innovador. Con el análisis de Big Data, las pequeñas y medianas empresas pueden obtener grandes volúmenes de datos semiestructurados para mejorar los diseños de sitios web e implementar sistemas efectivos de venta cruzada y recomendaciones personalizadas de productos. (S. Kamakhya, 2019)

La variedad es un problema que resulta significativo por la aportación masiva de información que se presenta en forma de datos estructurados (bases de datos, tablas, hojas de cálculo, etc.), semiestructurados (formatos tipo JSON o XML, etc.) y no estructurado (vídeos, imágenes, texto libre, etc.). (Cukier, 2010)

El problema de la variedad radica en la diversidad estructural de los datos, centrándose en la complejidad y multitud de fuentes, así como la combinación y diversidad de múltiples formatos, estructuras y lenguajes. (P. Shantanu, 2018)

Hasta el momento de publicación de este trabajo de tesis no se ha encontrado un trabajo de mapeo sistemático orientado al problema de la variedad utilizando herramientas o IDEs, frameworks o metodologías. En este sentido un estudio de mapeo sistemático proporciona una estructura del tipo de informes de investigación y resultados que se han publicado, categorizándolos y a menudo da un resumen visual, el "mapa", de sus resultados. (K. Petersen F. R., 2008)

La finalidad del mapeo sistemático es brindar un panorama general acerca del uso de herramientas, frameworks y metodologías que abordan el problema de la variedad en sistemas Big Data por medio de clasificaciones, representaciones estadísticas y visuales.

1.1 Contexto del Problema

En la actualidad Big Data es una de las tendencias principales dentro de la industria tecnológica. Si bien, el término “Big Data” es relativamente nuevo, el hecho de recopilar y almacenar grandes cantidades de información para su posterior análisis se viene realizando desde hace muchos años. El concepto cobró un mayor interés a principios de la década del 2000 cuando Doug Laney estructuró la definición de Big Data como las 3Vs: Volumen, Velocidad y Variedad. (Doug, 2001)

Volumen: El volumen se refiere a la cantidad de datos recopilados.

Velocidad. La velocidad se refiere a la rapidez en que los datos son creados, almacenados y procesados en tiempo real.

Variedad: La variedad se refiere a las formas, tipos y fuentes en las que se registran los datos.

En este contexto, la variedad es un problema en la actualidad, ya que los datos provienen de diferentes tipos de fuentes, haciendo que exista una diversidad en el formato y tipología en

ellos. Los datos pueden ser estructurados, semiestructurados o no estructurados, y sus fuentes pueden provenir de texto, imágenes, videos, audio, video, entre otros. (Doug, 2001)

De acuerdo a lo anterior actualmente las organizaciones se enfrentan a menudo a retos Big Data. Las empresas tienen acceso a una gran cantidad de información, pero no saben cómo obtener valor agregado de los datos, ya que la información aparece en su forma más cruda o en un formato semiestructurado o no estructurado. Una encuesta divulgada en (K. Navroop, 2019), concluye que más de la mitad de los líderes empresariales de hoy en día se dan cuenta de que no tienen acceso a los conocimientos que necesitan para analizar sus datos.

Las empresas se enfrentan a estos retos en ambientes en el que tienen la capacidad de almacenar cualquier cosa, que están generando datos como nunca antes en la historia y, sin embargo, tienen un verdadero desafío con el análisis de la información.

De acuerdo a (G. Kim, 2014) existe una gran variedad de aplicaciones de las técnicas de Big Data. Siempre que sea necesario extraer el conocimiento inmerso en grandes volúmenes de datos estructurados, semiestructurados o no estructurados, algunas de tantas pueden ser: Patrones de detección de fraudes, Patrones de Medios Sociales, Patrones de modelado y gestión de riesgo.

1.2 Descripción del problema

El desconocimiento del estado del arte de un área de investigación ocasiona la falta de información para proponer soluciones dentro del área en cuestión, en este sentido en el área de Big Data y específicamente dentro del problema de la variedad se identificó que no hay un estudio de mapeo sistemático que sirva como guía para abordar el análisis de la variedad estructural de los datos por medio de herramientas, frameworks o metodologías.

El análisis de la información de Big Data resulta muy útil para las empresas ya que proporciona respuestas a preguntas que suelen ser desconocidas, también ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. A su vez, esto lleva a cuestionarse ¿Cómo se pueden analizar o procesar los datos?, ¿Qué mecanismos o aplicaciones computacionales sirven para el análisis de la información?. Como se puede ver, es mucha la importancia del análisis de Big Data en las empresas. Por lo cual es necesario ampliar el conocimiento sobre estas nuevas tecnologías para saber cómo abordar la variedad de los datos en Big Data.

1.3 Objetivos

De acuerdo a la descripción del problema, en esta sección se presenta el objetivo general y los objetivos específicos de este trabajo de investigación.

1.3.1 Objetivo General

El objetivo general de este trabajo de investigación es evidenciar el estatus actual de como ha sido abordado el problema de la variedad en sistemas Big Data.

1.3.1 Objetivos Específicos

1. Evidenciar que Herramientas o IDEs han sido desarrollados y utilizados para abordar la variedad en sistemas Big Data.
2. Evidenciar que Frameworks han sido desarrollados y utilizados para abordar la variedad en sistemas Big Data.
3. Evidenciar que Metodologías han sido implementados para abordar la variedad en sistemas Big Data.

1.4 Justificación

Durante el transcurso de los años, hemos vivido la época de "la revolución de los datos". La expansión de internet y la tecnología han generado que los datos de ubicación, texto, video, imágenes, sensores, audio, o simplemente información estén disponibles en cualquier momento digitalmente y por ende ha traído consigo un crecimiento exponencial de nuevos tipos de datos, los cuales son difíciles de procesar con herramientas y frameworks previas al surgimiento de Big Data. Sin embargo, el mayor problema en el análisis de Big Data no es la gran cantidad de datos (Volumen), sino la dificultad de analizar los diferentes tipos de datos (Variedad).

En este contexto, un estudio de mapeo sistemático sobre el problema de la variedad en Big Data, y específicamente sobre el uso de herramientas, frameworks o metodologías para el análisis de los diferentes tipos de datos (Variedad); permite evidenciar las brechas, áreas de oportunidad y además brindar una guía para abordar el problema de la variedad.

1.5 Alcances y Limitaciones

1.5.1 Alcances

1. Se consideró como objeto de estudio artículos que trataron el problema de variedad en sistemas Big Data con: metodologías, frameworks y herramientas.

2. Se desarrolló un estudio de mapeo sistemático en el problema de la variedad en sistemas Big Data y de acuerdo a los resultados obtener un conjunto de herramientas, frameworks y metodologías para abordar el problema de la variedad por medio de su tipo de dato.

1.5.2 Limitaciones

1. Se accedió a las fuentes de datos del CENIDET por medio del CONRICYT: ACM digital library, ScienceDirect, IEEE Xplore, JSTOR y Springer Link.

2. Se consideraron artículos disponibles en texto completos.

3. No se realizó una revisión sistemática en su totalidad, pero si se hizo una revisión de cada uno de los trabajos finales seleccionados por el mapeo sistemático.

1.6 Resumen del capítulo.

El presente capítulo se expone el problema que dio origen al trabajo de investigación descrito en esta tesis; se definen los objetivos, la justificación, los alcances y limitaciones de la investigación.

Capítulo 2

Marco Conceptual

En este capítulo se presenta el marco conceptual, donde se explican los conceptos relacionados a Big Data y al estudio de mapeo sistemático, los conceptos presentados a continuación permitirán brindar un mejor entendimiento de los términos de esta investigación.

2.1 Estudio de Mapeo Sistemático

Los estudios de mapeo sistemático están diseñados para brindar un panorama de un área de investigación a través de la clasificación y conteo de contribuciones en relación a las categorías con el propósito de identificar áreas de investigación que no se ha abordado mucho y atacarlas. Este proceso implica buscar literatura acerca de áreas que han sido cubiertas y publicadas.

Deacuerdo a (K. Petersen F. R., 2008) define un estudio de mapeo sistemático como un estudio secundario que tiene como objetivo construir un esquema de clasificación y estructurar un campo de interés de la ingeniería de software. Para llevar a cabo el estudio sugiere un proceso de cinco pasos que son los siguientes: definir las preguntas de investigación, realizar la búsqueda de los estudios en las fuentes seleccionadas, selección de los estudios primarios, análisis de los resúmenes y extracción de palabras clave, extracción de datos, y por último mapear los estudios primarios seleccionados. Cada uno de los pasos del proceso tiene un resultado, siendo el mapeo sistemático el resultado final de los procesos: (1) Definición de Preguntas de Investigación, (2) Búsqueda, (3) Revisión de documentos, (4) Definición de palabras clave y (5) Extracción de datos y proceso de mapeo (Mapa sistemático).

2.2 Revisión Sistemática

De acuerdo a (Francisco, 2017) una revisión sistemática es un tipo de revisión de la literatura que recopila y analiza críticamente múltiples estudios o trabajos de investigación a través de un proceso sistemático.

Desde otro punto de vista según (C. Manterola, 2013) una revisión sistemática, es un artículo de “síntesis de la evidencia disponible”, en el que se realiza una revisión de aspectos

cuantitativos y cualitativos de estudios primarios, con el objetivo de resumir la información existente respecto de un tema en particular.

El objetivo principal de una SLR (Systematic Literature Review) es proporcionar un resumen exhaustivo de la literatura disponible pertinente a una o varias preguntas de investigación, una revisión sistemática es aquella en la que existe una búsqueda exhaustiva de estudios relevantes sobre un tema. Una vez identificados y obtenidos los estudios, los resultados son sintetizados de acuerdo con un método preestablecido y explícito. Esta forma de revisión da al lector una gran ventaja sobre otras revisiones: la posibilidad de replicarla y verificar si se llega a la misma conclusión. (Sáenz, 2018)

2.3 Principales diferencias entre un mapeo y un SLR

En (K. Petersen S. V., 2015) se revisa el concepto de mapeo sistemático y se explica la diferencia principal con una revisión sistemática y menciona que, aunque un estudio de mapeo sistemático y una revisión sistemática de la literatura comparten algunos puntos en común (por ejemplo, con respecto a la búsqueda y selección de estudios), son diferentes en términos de objetivos y, por lo tanto, enfoques para el análisis de datos.

Mientras que las revisiones sistemáticas tienen como objetivo sintetizar evidencia, considerando también la fuerza de la evidencia, los mapeos sistemáticos se preocupan principalmente por estructurar un área de investigación.

En (A. Barbara, 2010) mencionan que un mapeo sistemático emplea los mismos métodos que una revisión sistemática, pero por lo general es un estudio más sencillo de realizar ya que no tiene como objetivo analizar los resultados presentados en los artículos que conforman el mapeo y un mapeo sistemático puede ser la base para realizar posteriormente una revisión sistemática.

2.4 Big Data

Diferentes autores han dado diferentes definiciones de Big Data, por ejemplo, se dice que el término Big Data, en su sentido actual, fue introducido por primera vez en 2008 por el Grupo Gartner (Gartner, 2008), donde menciona que Big Data son activos de información de alto volumen, alta velocidad y/o gran variedad que demandan formas innovadoras y rentables de procesamiento de información que permiten una mejor comprensión, toma de decisiones y automatización de procesos.

En los trabajos (M. Khan, 2014), (G. Palak, 2015) y (Suresh, 2014) mencionan que "Big Data" se utiliza en general para describir la recopilación, procesamiento, análisis y visualización asociados con conjuntos de datos muy grandes.

En el trabajo de investigación (H. Hu, 2014) organizó las definiciones de Big Data en tres categorías comparativas, basadas en la comparación de características relacionales y Big Data, basadas en atributos desarrollados a partir de las características de Big Data y arquitectónicas que enfatizan la arquitectura informática y los elementos técnicos.

En la investigación (K. Stephen, 2013) Define al "Big Data" como la cantidad de datos más allá de la capacidad de la tecnología para almacenar, administrar y procesar de manera eficiente.

Si bien, el término “Big Data” es relativamente nuevo, el hecho de recopilar y almacenar grandes cantidades de información para su posterior análisis se viene realizando desde hace muchos años. El concepto cobró un mayor interés a principios del a década del 2000 cuando Doug Laney estructuró la definición de Big Data como las tres Vs: Volumen, Velocidad y Variedad. (C. Min, 2014)

2.4.1 Volumen

El volumen se refiere a la cantidad de datos recopilados. De acuerdo a (C. Min, 2014) el volumen significa que, con la generación y recopilación de grandes cantidades de datos, la escala de datos se vuelve cada vez más grande.

El volumen de acuerdo con (D. Rustem, 2017) es la cantidad de datos con la que se trabaja supera la capacidad de gestión del software habitual.

2.4.2 Velocidad

La velocidad se refiere a la rapidez en que los datos son creados, almacenados y procesados en tiempo real. De acuerdo a (C. Min, 2014) la velocidad significa que la puntualidad de los macro datos, específicamente, la recopilación y el análisis de datos, etc., debe llevarse a cabo de manera rápida y oportuna, de modo que se pueda utilizar al máximo el valor comercial de los grandes datos.

En el trabajo (D. Rustem, 2017) menciona que la rapidez es esencial a la hora de recibir, procesar y utilizar los datos en tiempo real.

2.4.3 Variedad

La variedad se refiere a las formas, tipos y fuentes en las que se registran los datos. El concepto de variedad fue propuesto por primera vez por Laney en 2001 con los siguientes elementos: resolución inconsistente, traducción "universal" basada en XML, Integración de Aplicaciones para Empresas (EAI), middleware de acceso a datos y Extracción, Transformación, Carga y Gestión (ETLM) de consultas distribuidas de metadatos. (Doug, 2001)

De acuerdo con (C. Min, 2014) la variedad indica los diversos tipos de datos, que incluyen datos semiestructurados y no estructurados como audio, video, página web y texto, así como datos estructurados tradicionales.

En la actualidad, el Instituto Nacional de Estándares Tecnología (NIST, por sus siglas en inglés) interpreta de una manera algo diferente la variedad, que describe la organización de los datos y si los datos son estructurados, semiestructurados o no estructurados. (Wilbur, 2019)

2.5 Variedad y tipos de datos

El trabajo (K. Stephen, 2013) menciona que "Big Data" originalmente se centraba en datos estructurados, pero la mayoría de los investigadores y profesionales se han dado cuenta de que la mayor parte de la información del mundo reside en información masiva y no estructurada, en gran parte en forma de texto e imágenes. La explosión de datos no ha ido acompañada de un nuevo medio de almacenamiento correspondiente.

El Big Data proviene de una gran variedad de fuentes y generalmente se presenta en tres tipos: estructurado, semiestructurado y no estructurado. (C. Eaton, 2012) Los datos estructurados insertan un almacén de datos ya etiquetado y ordenado fácilmente, pero los datos no estructurados son aleatorios y difíciles de analizar. Los datos semiestructurados no se ajustan a los campos fijos, pero contienen etiquetas para elementos de datos separados. (Singh, 2012)

De acuerdo con (Doug, 2001) los datos en el contexto de la variedad de Big Data pueden ser estructurados, semiestructurados o desestructurados, y sus fuentes pueden provenir de texto, imágenes, videos, audio, video, entre otros.

2.5.1 Datos Estructurados

Los datos estructurados tienen perfectamente definido la longitud, el formato y el tamaño de sus datos. Por lo general se almacenan en formato tabla, hojas de cálculo o en bases de datos relacionales. (S. Rolf, 2009)

Los datos estructurados básicamente son aquellos que se encuentran ordenados y organizados mediante una serie de filas y columnas bien definidas. Son los que se usan de manera habitual en la mayor parte de las bases de datos relacionales (RDBMS).

Dada su estructura ordenada, son los más fáciles de gestionar, tanto digital como manualmente. También, dada su alto grado de organización, permiten una mayor predictibilidad que otros tipos. (Keith, 2013)

2.5.2 Datos No estructurados

Los datos no estructurados se caracterizan por no tener un formato específico y por lo general se almacenan en múltiples formatos como documentos PDF o Word, correos electrónicos, ficheros multimedia de imagen, audio o video, etc. (S. Rolf, 2009)
De acuerdo con (Keith, 2013), estos datos no se pueden usar en una base de datos tradicional, ya que sería imposible ajustarlos a las filas y columnas estandarizadas ya que carecen de una estructura o un orden.

2.5.3 Datos Semiestructurados

Los datos semiestructurados son una mezcla de los datos estructurados y no estructurados, por lo general no presenta una estructura perfectamente definida como los datos estructurados, pero si presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones, y que en algunos casos están aceptados por convención, como por ejemplo los formatos HTML, XML o JSON. (S. Rolf, 2009)

De acuerdo a (Keith, 2013), son aquellos con un nivel medio de estructuración y rigidez organizativa. Se encuentran a medio camino entre los estructurados y los no estructurados. Un ejemplo válido sería un servidor local que almacenara todos los datos de correo electrónico y archivos adjuntos dentro de la base de datos. Tienen un cierto nivel de estructura, jerarquía y organización, aunque carecen de un esquema fijo.

2.5.4 Datos heterogéneos

Los datos heterogéneos son datos que pueden ser estructurados semiestructurados o no estructurados, a su almacenamiento y análisis en un corto periodo de tiempo, la mayoría de veces en tiempo real. (S. Rolf, 2009)

2.6 Herramienta

Las herramientas informáticas (tools, en inglés), son programas, aplicaciones o simplemente instrucciones usadas para efectuar otras tareas de modo más sencillo. En un sentido amplio del término, podemos decir que una herramienta es cualquier programa o instrucción que

facilita una tarea, pero también podríamos hablar del hardware o accesorios como herramientas. (B. Elizabeth. Sanders, 2010)

Las herramientas informáticas son un conjunto de instrumentos empleados para manejar información por medio de la computadora. El uso de estas herramientas, además de un conocimiento de la computadora requiere un conocimiento de las mismas en sus elementos, objetos que manejan y operaciones básicas; para sus aplicaciones se exige reconocer sus lógicas de uso, esquemas de organización y representación. (Manuel T. P., 2015)

2.7 IDEs

Deacuerdo a (RedHat, 2020) un entorno de desarrollo integrado (IDE) es un sistema de software para el diseño de aplicaciones que combina herramientas del desarrollador comunes en una sola interfaz gráfica de usuario (GUI). Generalmente, un IDE cuenta con las siguientes características:

- Editor de código fuente: editor de texto que ayuda a escribir el código de software.
- Automatización de compilación local: herramientas que automatizan tareas sencillas e iterativas como parte de la creación de una compilación local del software para su uso por parte del desarrollador.
- Depurador: programa que sirve para probar otros programas y mostrar la ubicación de un error en el código original de forma gráfica.

2.8 Framework

Un framework es un entorno de desarrollo completo, que suele facilitar herramientas tan indispensables como el compilador, el debugger y el editor de código. Pero también cuenta con un poderoso conjunto de bibliotecas, con funciones útiles ya previamente implementadas, que ahorran tiempo y esfuerzo al desarrollador y constituyen el núcleo del entorno. (B. Elizabeth. Sanders, 2010)

Un framework es una estructura de soporte al desarrollo de software en el cual aplicaciones de software pueden ser organizadas y desarrolladas. Un framework puede incluir programas de soporte, librerías, lenguaje de consultas, servicios, interfaces, u otras utilidades para

ayudar a desarrollar y unir los diferentes componentes de una aplicación de software. (Hua, 2006)

2.9 Metodología

El significado de metodología en sí, se refiere a los métodos de investigación que se siguen para alcanzar los objetivos en una ciencia o estudio, la metodología que se utilizará a lo largo de la investigación será la de estudio de caso de (Yin, 2012).

La metodología es la ciencia que nos enseña a dirigir determinado proceso de manera eficiente y eficaz para alcanzar los resultados deseados y tiene como objetivo darnos la estrategia a seguir en el proceso. (E. Manuel, 2004)

2.10 Método

Algunos autores definen el método como un procedimiento concreto que se emplea, de acuerdo con el objeto y con los fines de la investigación, para propiciar resultados coherentes. Es una serie de pasos sucesivos que conducen a una meta. (Manuel, Definicion De Terminos Basicos De Investigacion (Glosario), 2008)

El método es común a todas las ciencias, ya que se trata de un procedimiento riguroso formulado lógicamente, que permite adquirir un conjunto de conocimientos en forma sistemática y organizada. (Esther, 2014)

2.11 Modelo

Las acepciones del concepto de modelo son muy diversas. Puede considerarse al modelo, en términos generales, como representación de la realidad, explicación de un fenómeno, ideal digno de imitarse, paradigma, canon, patrón o guía de acción; idealización de la realidad; arquetipo, prototipo, uno entre una serie de objetos similares, un conjunto de elementos esenciales o los supuestos teóricos de un sistema social. (Caracheo, 2002)

Deacuerdo a (Manuel, 2013) define modelo como ejemplar o forma que uno propone y sigue en la ejecución de una obra artística o en otra cosa, ejemplar para ser imitado, representación en pequeño de una cosa, copia o réplica de un original, construcción o creación que sirve

para medir, explicar e interpretar los rasgos y significados de las actividades agrupadas en las diversas disciplinas.

2.12 Algoritmo

Un algoritmo es un conjunto reescrito de instrucciones o reglas bien definidas, ordenadas y finitas que permite realizar una actividad mediante pasos sucesivos que no generen dudas a quien lo ejecute. Dados un estado inicial y una entrada, siguiendo los pasos sucesivos se llega a un estado final y se obtiene una solución. Los algoritmos son objeto de estudio de la algoritmia. (Vicente, 2015)

Un algoritmo es una secuencia de operaciones detalladas y no ambiguas, que, al ejecutarse paso a paso, conducen a la solución de un problema. En otras palabras, es un conjunto de reglas para resolver una cierta clase de problema. (Luis, 2020)

2.13 Clúster

Un clúster es un conjunto de computadoras que utilizan componentes comunes y actúan como si se tratase de un solo sistema u ordenador. (Victoria, 2009)

Un clúster de computadoras de acuerdo con (UNAM, 2001), es un sistema de procesamiento paralelo o distribuido. Consta de un conjunto de computadoras independientes, interconectadas entre sí, de tal manera que funcionan como un solo recurso computacional. A cada uno de los elementos del clúster se le conoce como nodo. Estos son aparatos o torres que pueden tener uno o varios procesadores, memoria RAM, interfaces de red, dispositivos de entrada y salida, y sistema operativo.

2.14 Resumen del capítulo

En este capítulo se mostraron los principales conceptos para entender el contexto de este trabajo de tesis. Se describe de manera general lo que es un estudio de mapeo sistemático y su principal diferencia con una revisión sistemática de la literatura, también se presentan los conceptos principales para abordar el concepto de Big Data. Todos los conceptos presentados son de relevancia para entender los siguientes capítulos. El siguiente capítulo presenta la metodología para realizar el estudio de mapeo sistemático.

Capítulo 3

Estudio del Arte

En este capítulo se presenta un trabajo antecedente (una investigación realizada en CENIDET) y 25 trabajos relacionados que sirvieron de base para este trabajo de investigación.

3.1 Antecedentes

3.1.1 Mapeo sistemático del reconocimiento del habla, proceso del lenguaje natural y uso de ontologías para identificar el dominio del problema y los requerimientos de solución

En esta tesis de maestría (Santiago, 2019), presenta un mapeo sistemático con el fin de reunir trabajos publicados acerca de distintas técnicas de reconocimiento del habla, procesamiento de lenguaje natural y uso de ontologías en la elicitación de requerimientos, dentro del periodo 2010-2018.

Como principal aportación de este trabajo de investigación es un mapeo sistemático que muestra una clasificación de las técnicas más utilizadas en el flujo de trabajo del Procesamiento de Lenguaje Natural. Las principales técnicas para el procesamiento del lenguaje natural fueron: Reconocimiento del habla, pre-procesamiento de texto, análisis de texto, análisis de datos exploratorios, representación de texto, ingeniería de características y la extracción de patrones. Además, se realizó una clasificación de cada técnica junto con sus fases y descripciones.

3.2 Ampliación del estudio del arte

3.2.1 Systematic Mapping Studies in Software Engineering

Este artículo de investigación (K. Petersen F. R., 2008) desarrolla un mapeo sistemático en el área ingeniería de software con el fin de construir un esquema de clasificación y estructurar un campo de interés. El análisis de los resultados se centra en las frecuencias de publicaciones para categorías dentro del esquema.

El mapeo sistemático propuesto en este trabajo consistió en las siguientes etapas:

1. Definición de preguntas de investigación (ámbito de investigación)
2. Realizar búsqueda de estudios primarios (todos los documentos)
3. Selección de documentos para inclusión y exclusión (documentos relevantes)
4. Palabras clave de los resúmenes (esquema de clasificación)
5. Extracción de datos y mapeo de estudios (Mapeo sistemático)

Concluye que los mapas sistemáticos y las revisiones son diferentes en términos de objetivos, amplitud, problemas de validez e implicaciones.

3.2.2 Descubrimiento de Conocimiento en Big Data: Estudio de Mapeo Sistémico

Este artículo de investigación (F. T. Luis, 2015) realizó un estudio de mapeo sistémico, enfocado al Descubrimiento de Conocimiento (KDD) en Big Data, la metodología empleada fue la siguiente:

- Definición de las Preguntas de Investigación
- Selección de Fuentes
- Conducción de la Búsqueda
- Selección de Estudios
- Extracción y Síntesis de Datos.

Algunas preguntas de investigación fueron las siguientes:

¿Qué tendencias existen sobre la implementación de KDD en entornos Big Data?, ¿Qué proceso KDD se aborda principalmente?, ¿Cuáles son los principales retos relacionados con la implementación de KDD en entornos de Big Data?, ¿Qué proceso KDD se aborda principalmente? El resultado final de este trabajo da una conclusión acerca de cómo la Ingeniería de Software es necesaria para abordar las soluciones de Big Data Analytics.

3.2.3 A Systematic Mapping Study in Microservice Architecture

Este artículo de investigación (A. Nuha, 2016) realizó un estudio de mapeo sistemático de las arquitecturas de microservicios y su implementación, centrándose en la identificación de los problemas arquitectónicos, vistas arquitectónicas y los atributos de calidad.

La estructura que sigue el estudio del mapeo sistemático fue el siguiente:

- Definición de las Preguntas de investigación:
  - ¿A qué retos arquitectónicos se enfrentan los sistemas de microservicios?
  - ¿Qué diagramas / vistas de arquitectura se utilizan para representar arquitecturas de microservicios?
  - ¿Qué atributos de calidad relacionados con los microservicios se presentan en la literatura?
- Estrategia de búsqueda.
- Selección de estudios primarios.
- Palabras clave y clasificación.
- Criterios de Inclusión y exclusión.
- Estrategia de extracción de datos y evaluación de la calidad.

El resultado final de este trabajo concluye que los retos arquitectónicos en los sistemas de microservicios son: la ccomunicación/integración, descubrimiento de servicios, rendimiento, tolerancia a fallos, seguridad, seguimiento y registro. Proponen una vista / lenguaje de modelado integral basada en UML y por último describen 15 atributos de calidad para la validación de microservicios.

3.2.4 Research on Big Data - A systematic mapping study

Este artículo de investigación (A. Jacky, 2017) examina cómo los investigadores comprenden el concepto de Big Data. Definiendo una serie de preguntas:

- ¿Cuántos trabajos de investigación se producen?
- ¿Cuál es la tendencia anual de las publicaciones?
- ¿Cuáles son los temas principales en la investigación de Big Data?
- ¿Cuáles son los temas de Big Data más investigados?
- ¿Por qué se realiza la investigación?
- ¿Qué produce la investigación de Big Data?
- ¿Quiénes son los autores activos?
- ¿Qué revistas incluyen artículos sobre Big Data?
- ¿Cuáles son las disciplinas activas?

La investigación se realizó mediante un estudio de mapeo sistemático de publicaciones basadas en un rango de 10 años. Los resultados obtenidos demostraron que la comunidad de investigación ha realizado importantes contribuciones, como lo demuestra el continuo aumento en el número de publicaciones que tratan sobre Big Data.

El proceso del estudio sistemático fue de la siguiente manera:

1. Definición de las preguntas de investigación
2. Alcance de la revisión
3. Búsqueda de artículos
4. Selección de artículos
5. Procesos de Inclusión y Exclusión de artículos
6. Revisión del Abstract por palabras Clave
7. Esquema de Clasificación
8. Extracción de la información
9. Mapeo Sistemático

Los principales resultados obtenidos fueron:

(i) Hay un crecimiento significativo de artículos de investigación sobre Big Data desde 2013.

(ii) Existe una diversidad de intereses por parte de los investigadores en temas como los objetivos, los artefactos producidos, los criterios de calidad utilizados y los usos y aplicaciones de Big Data.

(iii) Big Data Research se centra principalmente en tres técnicas, es decir, agrupación, clasificación y predicción.

3.2.5 A Generic Framework for Concept-Based Exploration of Semi-Structured Software Engineering Data

De acuerdo al artículo de (Gillian, 2015) presenta una investigación y propone el desarrollo de un framework genérico sobre la exploración basada en conceptos de conjuntos de datos semiestructurados de ingeniería de software por medio de una combinación de etiquetas y redes conceptuales. Para ello se realizó un estudio sistemático con las siguientes características:

- Definición de las preguntas de Investigación.
  - ¿Cuál es una estructura de datos adecuada para una variedad de tareas de análisis de ingeniería de software?
  - ¿Cómo podemos hacer que los algoritmos de red sean escalables para grandes conjuntos de datos?
  - ¿Cómo se pueden construir redes automáticamente a partir de una variedad de conjuntos de datos?
  - ¿Qué es una interfaz adecuada y escalable para una variedad de tareas de análisis de datos de ingeniería de software?

- Metodología de investigación: en esta investigación se siguió un enfoque de "construir y probar". Se construyó un framework para la exploración de datos semiestructurados de ingeniería de software a partir de una variedad de aplicaciones de referencia.

- Enfoque técnico:
  - Construcción conceptual.
  - Construcción de Interfaz.
  - Soporte para la navegación

- Implementación del prototipo.

3.2.6 Tactical Big Data Analytics: Challenges, Use Cases, and Solutions

Este artículo de investigación (S. Onur, 2013) clasifica los desafíos tácticos del análisis de Big Data con respecto a los datos y presentan un framework de solución integral motivado por los casos de uso relevantes.

Mencionan los retos más importantes para el análisis táctico de Big Data:

- Fuentes de datos heterogéneos: La mayoría de los datos del DoD no están estructurados, como señales, texto, imágenes y video con diferentes estandarizaciones.
- Datos inciertos / incompletos / ruidosos: Las incertidumbres pueden surgir de varias inexactitudes y deberían estar representadas en la estructura de datos.
- Metas orientadas a la misión: Las metas en el problema táctico del Big Data son impulsadas por estrictas necesidades de las aplicaciones comerciales.
- Exigentes requisitos de seguridad: El hecho de que los datos militares se encuentren en el mismo entorno virtual que otras ofertas comerciales puede no cumplir con los estrictos requisitos de seguridad del DoD (como la protección contra el robo de datos y los ataques de corrupción).

Mencionan como algunas herramientas pueden tratar con estos problemas, como lo son:

- Hadoop Core
- MapReduce
- Apache Pig
- HBase
- Hive
- Storm

Llegando a la conclusión de que las herramientas de mashup han tenido éxito en apoyar el desarrollo de aplicaciones para el internet de las cosas. Por otra parte, la integración de las herramientas existentes de mashup con las herramientas de análisis de Big Data puede permitir el desarrollo continuo de aplicaciones sofisticadas de alto impacto.

3.2.7 Variety Management for Big Data

Este artículo de investigación (M. Wolfgang, 2018) definió y analizó los diferentes tipos y fuentes de variedad de datos en Big Data. Introduce el concepto de metadatos semánticos como base para describir y gestionar la variedad en el contexto de Big Data.

Utiliza los metadatos semánticos para capturar las fuentes de la variedad en Big Data desde 5 perspectivas las cuales son:

1. Variación estructural (por estructura de almacenamiento, formato de datos o variación semántica),
2. Variaciones en la granularidad (ya sea por agregación a lo largo del eje temporal),
3. Fuentes de datos heterogéneas,
4. Grados de calidad e integridad, y
5. Diferencias en los datos (pre procesamiento)

Se hace mención de que las ontologías pueden ayudar a descubrir, navegar, explorar e interpretar una variedad de datos heterogéneos. Las ontologías y los catálogos completos de metadatos pueden simplificar la interpretación, mejorar la calidad de los datos y simplificar la integración de múltiples conjuntos de datos.

3.2.8 From Text to XML by Structural Information Extraction

Este artículo de investigación (Y. Piao, 2015) muestra la extracción de la estructura del texto libre y su conversión al formato XML, con un algoritmo basado en CRF (Conditional Random Field model) SIECRF (Structure Information Extraction model based on CRF).
Muestra varios casos de prueba con un enorme volumen de XML semiestructurado y de texto libre no estructurado por medio de la recuperación de información de la red.
Este artículo estudia las cuestiones clave relacionadas con la extracción de la estructura del texto basada en CRF y se introduce el algoritmo correspondiente SIECRF, incluidas las formas de construir funciones de características para el modelo considerando la estructura característica de XML.
Las principales características presentadas son:
La extracción de información tiene aproximadamente dos tipos de modelos

- Basado en reglas: resume, organiza y reconoce la entidad y su relación, luego almacena estos conocimientos en una cierta forma de descripción
- Basado en estadísticas: tienen una base matemática sólida, lo que lo hace más práctico y conveniente para manejar textos masivos y sin estructuras.

3.2.9 Adaptive System for Handling Variety in Big Text

Este artículo de investigación (P. Shantanu, 2018) propone un sistema para procesar la variedad de texto con diferentes formatos, tamaños, idiomas y contextos a la perfección, abarcando texto generado a través del ecosistema de Big Data.

Los textos trabajados son obtenidos de diversas fuentes como bases de datos, tablas, hojas, páginas web y las redes sociales, generando variedad en distintos formatos.

Características:

- Este trabajo introduce un método para entrenar ingresando el clasificador con texto multilingüe.
- Los pasos principales en el sistema son el pre-procesamiento y la construcción de modelos de texto.

3.2.10 A general perspective of Big Data: applications, tools, challenges and trends

Este artículo de investigación (R. Lisbeth, 2015) realiza una revisión de la literatura proporcionando una visión general sobre el estado del arte en aplicaciones Big Data de los últimos 4 años, esto con el fin de identificar los principales desafíos, áreas de aplicación, herramientas y tendencias emergentes de Big Data. La clasificación que realizó lo hizo dependiendo de los problemas de Big Data, quedando de la siguiente manera:

- Captura, almacenamiento, búsqueda, análisis y visualización de datos

También se lleva a cabo una clasificación de las herramientas de Big Data, quedando de la siguiente manera:

- Herramientas de Big Data basadas en el análisis por lotes:
  - Google MapReduce
  - Apache Hadoop, Mahout, Storm, S4, Spark
  - Microsoft Dryad
  - Stream analysis
  - MOA
- Herramientas de Big Data basadas en el análisis interactivo:
  - Apache Drill, SpagoBI y D3.js

3.2.11 Big Data Validation Case Study

Este artículo de investigación (X. Chunli, 2017) realiza un estudio para la dimensión de la calidad de los datos, el proceso de validación de datos y las herramientas de Big Data, para ello realizan una serie de estudios de casos de calidad de Big Data. El estudio muestra las herramientas de validación, el proceso de validación y el resultado de la validación de datos. Para el proceso de validación de Big Data realizó:

- Recopilación de datos
- Limpieza de datos
- Transformación de datos
- Carga de datos
- Validación de datos
- Análisis de datos e informe

Para el caso de estudio se llevó a cabo de la siguiente manera:

- Descripción del caso de estudio
- Herramientas de validación de Big Data
- Resultados del caso de estudio

Por último, se establecen algunos criterios de validación de Big Data y se seleccionan algunas herramientas de validación de datos para realizar un estudio sobre datos de sensores meteorológicos. Las funciones, características y limitaciones de las herramientas se analizan a detalle. Se verifican las dimensiones de calidad de datos seleccionadas y se presentan los resultados. Se menciona que algunos problemas de calidad de datos no se han resuelto y deberían estudiarse en el futuro.

3.2.12 Big Data: framework and issues

Este artículo de investigación (H. Lamyae, 2016) presenta la tecnología de Big Data junto con su importancia en el mundo moderno, destaca sus campos clave y problemas de mayor importancia, también analiza en detalle las herramientas de Big Data (Hadoop y Spark) y las compara de acuerdo con varios criterios. El artículo aporta con una solución a los problemas de almacenamiento y seguridad de Big Data.

Describe los campos claves de Big Data, ya que son importantes para las organizaciones de cualquier industria, por ejemplo:

- Comercio: el análisis de Big Data proporcionaría una mejor comprensión de los comportamientos y preferencias de los clientes, ayudando a probar la eficiencia de las estrategias comerciales, mejora el rendimiento y optimiza la distribución y la logística.

- Los grupos de mercadotecnia y publicidad: utilizando de manera correcta las redes sociales para el uso de la promoción y anticipar el deseo o la demanda del consumidor.
- Atención médica: ayudando a comprender mejor la evolución de una enfermedad, tomar mejores decisiones médicas, personalizar la medicina, etc.
- Banca: obteniendo la mayor cantidad de datos en forma de texto y cifras, y así predecir sucesos que puedan beneficiar a la economía.

Algunas herramientas que ayudan para el análisis de toda esta información son:

- Hadoop	- Hive	- Ambari
- HDFS	- Mahout	- Apache Spark:
- MapReduce	- Oozie	- Spark Core (API)
- HBase	- Sqoop	- Spark clustering
- Pig	- ZooKeeper	- Spark stack

Este artículo también describe los desafíos y problemas de Big Data, pero también describe como el análisis de Big Data ayudaría a las organizaciones empresariales a avanzar hacia esta tecnología para aumentar el valor. Sin embargo, las tecnologías tradicionales no son capaces de manejar los desafíos de Big Data, requieren tecnologías potentes y métodos avanzados para garantizar el rendimiento, la confiabilidad de los resultados, la disponibilidad de datos y la escalabilidad.

3.2.13 Big Data: Issues, Challenges, and Techniques in Business Intelligence

Este artículo de investigación (M. Ahmad, 2018) identificó los problemas y desafíos más relevantes relacionados con Big Data y señala una comparación exhaustiva de varias técnicas para manejar el problema de Big Data.

La inteligencia de negocios cubre una serie de herramientas, aplicaciones y métodos que ayudan a las empresas a recopilar datos de fuentes internas y externas, prepararlo para el análisis, crear y ejecutar consultas para obtener información valiosa de los datos, generar informes y gráficos, de modo que los resultados analíticos generados ayudarán a las organizaciones a tomar decisiones precisas y rápidas.

Se han sugerido y empleado dos métodos para el procesamiento de Big Data: procesamiento de datos almacenados por lotes y procesamiento de flujo de datos en tiempo real.

Las principales características que muestran son:

- Incluir métodos de análisis estadístico o cualitativo para la inteligencia de negocios.
- Minería y análisis de datos
- Modelado predictivo
- Análisis de Big Data y análisis de texto para una toma de decisiones efectiva

Sugieren dos métodos para el procesamiento de Big Data: procesamiento de datos almacenados por lotes y procesamiento de flujo de datos en tiempo real.

- Utilizar Hadoop como plataforma de procesamiento de grandes datos.
- Utilizar Apache Spark proporcionando un enfoque unido para administrar los requisitos de procesamiento de Big Data.

3.2.14 Efforts toward Research and Development on Inconsistencies and Analytical tools of Big Data

Este artículo de investigación (K. Ravindra, 2015) se centra en el análisis de Big Data, así como también en sus distintas dimensiones. También realiza una comparación entre las herramientas analíticas de Big Data de licencia y las herramientas analíticas de código abierto, seleccionando el tipo correcto de herramientas para el análisis de Big Data.

La complicación en el análisis de Big Data es por causa de las inconsistencias en los datos. Las inconsistencias se identificaron y diferenciaron en los niveles de datos, información, conocimiento y meta conocimiento. Algunas inconsistencias de las cuales aborda este artículo son:

- Inconsistencias temporales: cuando los conjuntos de datos contienen un atributo temporal y suelen ser conflictivas.
- Inconsistencias espaciales: ocurren debido a violaciones en un conjunto de datos que incluyen propiedades geométricas (ubicación, forma).
- Inconsistencias en el texto: la condición en la que dos textos se refieren al mismo evento o entidad, lleva a generar inconsistencias.
- Inconsistencias funcionales de dependencia: las violaciones de tales dependencias funcionales o dependencias funcionales condicionales darán como resultado inconsistencias en los datos e información.

Las comparaciones entre las herramientas analíticas de Big Data fueron:

- Pentaho
- TerraEchos
- Cognos
- Attivio
- Google BigQuery
- Netezza
- Apache Hadoop
- Zettaset
- HPCC Systems
- Dremel
- MarketGreenplum HD
- HortonWorks
- ParAccel
- GridGrain

En este artículo se describieron los principales problemas y desafíos multidimensionales en Big Data. Se identificaron las inconsistencias y se describieron varios tipos. Finalmente, se realizó una comparación sobre las herramientas analíticas de Big Data. El análisis de herramientas de Big Data tanto de licencia como de código abierto, se puede decir que cada herramienta depende principalmente del uso y la necesidad del individuo o de la empresa que lo use. Los problemas críticos en las herramientas de código abierto son modificaciones y desactualización.

3.2.15 Evolution of Spark Framework for simplifying Big Data Analytics

Este artículo de investigación (Subhash, 2016) analiza cómo se puede usar el framework Spark para Big Data Analytics. También se menciona que el procesamiento de los datos intensivo de la CPU se puede resolver rápidamente en Spark.

Este artículo se enfoca en el ecosistema de Spark, también compara el framework Spark con el framework de Flink. Se menciona que Spark se puede integrar con la herramienta Hadoop para llevar a cabo el procesamiento intensivo de los datos. Este modelo resultante tiene varias capas dentro del procesamiento de los datos como son:

- MLBase
- GraphX
- Spark Streaming
- Apache Kafka
- Spark SQL

Por último, se hace la comparación de Spark con Flink donde se menciona que Flink es un nuevo framework para manejar el análisis de Big Data por las siguientes características:

a) Optimizador
b) Manejo de Algoritmos
c) Utiliza la Serialización y deserialización en flujo de datos
d) Gestión eficiente de la memoria

Como conclusión se menciona que el framework Spark se usa ampliamente para Big Data Analytics actualmente, pero Flink podría ser la nueva opción para el manejo del procesamiento y análisis de Big Data.

3.2.16 Big Data and the SP Theory of Intelligence

Este artículo de investigación (Gerard, 2014) trata sobre cómo la teoría de la inteligencia de la simplicidad y potencia (SP, por sus siglas en inglés) en construcción de la máquina de SP puede aplicarse a la gestión y análisis de Big Data. El sistema puede descubrir estructuras "naturales" en Big Data, y tiene fortalezas en la interpretación de datos, incluyendo cosas como el reconocimiento de patrones, el procesamiento del lenguaje natural, varios tipos de razonamiento y más.

Características:

- Tiene fortalezas en el aprendizaje no supervisado o en el descubrimiento de estructuras en los datos, en el reconocimiento de patrones, en el análisis y la producción de lenguaje natural.
- Se presta para el análisis de datos de transmisión, ayudando a superar el problema de la velocidad en Big Data
- En el sistema SP, todo tipo de conocimiento se representa con patrones: matrices de símbolos atómicos en una o dos dimensiones.

3.2.17 A Comparative Study to Classify Big Data Using Fuzzy Techniques

Este artículo de investigación (Soha, 2016) tiene por objetivo implementar técnicas de clasificación para optimizar el framework de reducción de mapas usando métodos difusos.

El método aplicado para la técnica difusa es el dato *k* más cercano, y para las técnicas no difusas utilizan tanto una máquina de vectores de soporte. El uso del paradigma de reducción de mapas lo aplican para poder procesar grandes datos.

También implementan un sistema integrado, utilizando una máquina de vectores de soporte con la etiqueta difusa y la etiqueta difusa gaussiana.

Implementan tres técnicas de clasificación algorítmica; utilizan modelos difusos esto con el fin de construir un clasificador individual para cada grupo de datos, el procesamiento de los datos es mediante modelos y algoritmos matemáticos.

También utilizan los modelos de MapReduce para el procesamiento de datos en escala horizontal agrupados.

Utilizan la función gaussiana de la siguiente manera: La definición de la función de Fuzzy es la siguiente: $\mu A: X \rightarrow [0,1]$, donde a cada elemento en el vector *X* se le asigna un valor entre 0 y 1 y luego se le asigna. El valor asignado se denomina valor de miembro.

Para concluir, en este artículo presentan un estudio comparativo de clasificación en Big Data, utilizando tres algoritmos de clasificación; el dato *k* más cercano, el dato difuso *k* más cercano y la máquina de vectores de soporte que utiliza MapReduce. También proponen modelos matemáticos y algoritmos para el procesamiento difuso de datos.

3.2.18 A Big Data-as-a-Service Framework: State-of-the-Art and Perspectives

Este artículo de investigación (W. Xiaokang, 2018) su objetivo principal fue revisar el estado del arte de Big Data desde los aspectos de organización y representación, limpieza y reducción, integración y procesamiento, seguridad y privacidad, análisis y aplicaciones de Big Data.

Un aporte relevante de este artículo es la presentación de un framework llamado Big Data-as-a-Service. El framework que presentan consta de tres planos: el plano de detección, el plano de nube y el plano de aplicación, esto con el fin de abordar sistemáticamente los desafíos de organización, limpieza, integración y procesamiento.

La representación y reducción de datos que presentan es mediante las siguientes fases:

- Organización y representación de datos
  - Representación gráfica.
  - Representación ontológica.
  - Representación difusa.
  - Representación Tensor.
- Limpieza y reducción de datos
  - Análisis de componentes principales (PCA).
  - Análisis de componentes principales del núcleo (KPCA).
  - Descomposición de valor singular (SVD).
  - Análisis de componentes independientes (ICA).
  - Análisis discriminante lineal (LDA).
- Integración y procesamiento de datos
  - Computación en la nube.

Para concluir, en este artículo presenta un nuevo framework de Big Data como servicio para representar, reducir, integrar y procesar Big Data.

3.2.19 Quantifying Volume, Velocity, and Variety to Support (Big) Data-Intensive Application Development

Este artículo de investigación (D. Rustem, 2017) tiene como objetivo capturar y modelar las 'tres V' de Big Data para proporcionar información útil sobre el proceso general de los datos a partir de los atributos V de Big data. Proponen un framework para proporcionar una estimación de las métricas de los V-atributos mediante la evaluación de un modelo de rendimiento generado a partir del proceso de datos.

También presentan como se debe de abordar y analizar Big Data para una nueva generación de sistemas de software centrados en datos, como lo son las Aplicaciones Intensivas en Datos (DIA): que sirven para extraer valor comercial de Big Data.

Para ello proponen un enfoque para el desarrollo de software, en el que las 'tres V' de Big Data, se toman en consideración para estimar las demandas generales de recursos de aplicaciones basadas en las características de sus componentes de flujo de trabajo individuales, en tiempo de diseño.

El enfoque propuesto define un proceso de desarrollo iterativo basado en las V para los DIA, que abarca las fases de diseño, desarrollo e implementación del ciclo de vida del software.

Por lo tanto, la contribución principal de este artículo son las siguientes:

i. Caracterizar cuantitativamente y definir los V-atributos para un DIA a través de métricas específicas
ii. Especificar un proceso de desarrollo de DIA impulsado por las V-métricas, siguiendo enfoques y metodologías de ingeniería de software
iii. Definir una técnica de modelado y evaluación para procesos de Big Data en DIA, combinando la reducción del flujo de trabajo y las redes de colas para apoyar a los arquitectos de DIA.

La novedad de este trabajo en comparación con los existentes puede identificarse no solo en el dominio de Big Data y DIA, sino también en el enfoque de métricas y técnicas adoptadas que incluyen la reducción del flujo de trabajo y las redes de colas para el análisis de datos.

Para concluir, este artículo propone un enfoque, que tiene en cuenta los atributos de Big Data, como el volumen, la velocidad y la variedad. También presenta un modelado de las V-métricas para identificar el número exacto de nodos para manejar las tareas de procesamiento de datos.

3.2.20 BIG Data and Methodology-A review

Este artículo de investigación (K. Manjit, 2013) muestra una vista general del Big Data desde su definición, parámetros de las características de Big Data, su evolución, así como también las tecnologías para gestionar y procesar los diferentes tipos de datos por medio de las siguientes tecnologías:

- Hadoop
- MapReduce
- HDFS
- HBase
- Pig
- Hive
- Sqoop
- Avro
- Oozie
- Chukwa
- Flume
- Zookeepe

Las aportaciones más destacables de este artículo son las descripciones de las técnicas y tecnologías para el análisis de los datos.

3.2.21 Cloud resourcemanagement using 3Vs of Internet of Big Data streams

Este artículo de investigación (N. Kaur, 2019) propone un método que predice las características de los datos de la transmisión del Internet del Big Data (IoBd) en términos de volumen, velocidad y variedad (3V).

El proceso llevado es mediante valores pronosticados que se expresan en términos de Caracterización de Stream (CoSt). Por último, se asigna un clúster a la secuencia IoBd en función de su CoSt. Para ello el método propuesto en este artículo se utiliza el filtro de Kalman para la predicción de 3V junto con Mapas de Auto Organización (SOM) para la formación de grupos de datos.

El método propuesto utiliza dos módulos que interactúan entre sí para lograr los objetivos requeridos. El primer módulo, denominado Workcaster Forecaster (Wo-Fo), examina el flujo entrante de IoBd y pronostica la carga de trabajo que se espera que llegue durante el próximo intervalo de tiempo. Wo-Fo expresa la carga de trabajo prevista en términos de un triplete llamado Caracterización de Stream (CoSt). El segundo módulo, denominado Resource

Manager (RM), crea dinámicamente grupos de recursos en la nube basados en CoSt. La solicitud entrante utiliza los recursos asignados para producir la salida deseada.

En conclusión, este artículo propone un método que programa las transmisiones de IoBd a través de la nube en tiempo real. El método propuesto también mejora la utilización de los recursos, la disponibilidad y el tiempo de respuesta de los recursos en la nube. Por lo tanto, la metodología propuesta es una forma eficiente de procesar flujos de IoBd a través de la nube.

3.2.22 Big Data Reduction Methods: A Survey

Este artículo de investigación (H. Muhammad, 2016) presenta una revisión de los métodos que se utilizan para la reducción de Big Data. También presenta una discusión taxonómica detallada de los métodos de reducción de Big Data, incluyendo la teoría de redes, la compresión de Big Data, la reducción de dimensiones, la eliminación de redundancia, la minería de datos y los métodos de aprendizaje automático.

Las principales contribuciones de este artículo son:

- Presenta una revisión exhaustiva de la literatura y la clasificación de los métodos de reducción de Big Data.
- Esquemas propuestos para la reducción de Big Data.

Los métodos de reducción de Big Data que presentan son:

- Teoría de la red.
- Compresión.
- Eliminación de redundancia.
- Pre procesamiento de datos.
- Reducción de dimensiones.
- Minería de datos y aprendizaje automático (DM y ML)

En conclusión, se muestran métodos para abordar el problema de la reducción de Big Data. Y se menciona que no existe un método que pueda manejar el problema de la complejidad de Big Data de forma individual al considerar las 6V. En general, los métodos de reducción de datos basados en compresión son convenientes para reducir el volumen.

3.2.23 Metodología para el modelamiento de datos basado en Big Data, enfocados al consumo de tráfico (voz-datos) generado por los clientes

Este artículo de investigación (Sebastian, 2016) habla de Big Data y de cómo esta nueva tecnología ayuda a las organizaciones a tomar mejores decisiones, realizando un análisis del consumo (Voz - Datos) que estos generan al llamar por su teléfono.

Este artículo busca generar estrategias comerciales en tiempo real para el tratamiento y almacenamiento de datos, en entornos de gran volumen, variedad de orígenes y en los que la velocidad de respuesta es crítica.

Proponen un diseño de una metodología para el análisis de información a través del Big Data, enfocados al consumo de tráfico (Voz-Datos) generado por los clientes de una organización del sector de las telecomunicaciones, se puede identificar el valor ganado que se podrá utilizar posteriormente con herramientas orientadas en Big Data. Por lo cual se podrá tomar una decisión acertada.

Este articulo concluye con la importancia de la metodología propuesta y hace mención que es de vital importancia para las organizaciones, ya que dará una pauta de cómo abordar y manipular grandes cantidades de datos, y así poder tomar una toma de decisiones más correcta.

3.2.24 A Big Data methodology for categorising technical support requests using Hadoop and Mahout

Este artículo de investigación (D. Arantxa, 2014) propone una solución de prueba de concepto (PoC) de extremo a extremo utilizando el modelo de programación Hadoop, Mahout y con ayuda del Big Data Analytics para categorizar llamadas de soporte similares para grandes conjuntos de datos de soporte técnico.

Las contribuciones de este trabajo son las siguientes:

- Se describe una solución de para procesar, analizar y clasificar llamadas de soporte técnico. La solución propuesta utiliza la plataforma de procesamiento de datos distribuidos de Hadoop y las técnicas de agrupamiento en paralelo utilizando la biblioteca Mahout.

- En segundo lugar, se realiza una evaluación del rendimiento y la precisión de los algoritmos de agrupamiento en paralelo para analizar un conjunto de datos distribuidos utilizando un conjunto de datos de soporte técnico del mundo real.

La solución propuesta proporciona una solución de extremo a extremo para realizar análisis a gran escala de datos de soporte técnico utilizando la plataforma Hadoop de código abierto, componentes del ecosistema Hadoop como HBase y Hive y algoritmos de agrupamiento de la biblioteca extendida de Mahout.

Se concluye que la investigación presentada en el artículo presenta una solución completa de código abierto para el procesamiento y la categorización de llamadas de servicio similares dentro de grandes conjuntos de datos de soporte técnico para permitir la identificación de llamadas similares con potencial para una resolución más rápida.

3.2.25 Discusión de trabajos relacionados

En la Tabla 1. Comparación de trabajos relacionados se muestra una comparativa entre los trabajos que fueron estudiados para obtener un panorama acerca del análisis general de Big Data.

Como se puede observar los primeros cinco trabajos relacionados son artículos con relación a un estudio de mapeo sistemático, cabe resaltar el estudio de (K. Petersen F. R., 2008) ya que sirvió como guía fundamental para realizar un estudio de mapeo sistemático en varios trabajos de investigación, dicho estudio muestra los procesos y etapas que se tienen que realizar para la elaboración de un mapeo sistemático en la ingeniería de software, pero siguiendo el proceso del mapeo sistemático es posible emplearlo en cualquier estudio de investigación.

El trabajo de (M. Wolfgang, 2018) tiene un enfoque de investigación de la característica de la variedad, se centra en el análisis de los diferentes tipos de datos llamados “homogéneos” y los caracteriza desde varias perspectivas, esto con el fin simplificar la integración de múltiples conjuntos de datos por medio de algoritmos y ontologías.

El trabajo de (Y. Piao, 2015) y (P. Shantanu, 2018) tienen un enfoque relacionado al análisis del texto ya que el texto es un tipo de dato no estructurado. El objetivo del primer trabajo

consiste en extraer los datos más relevantes de un documento para facilitar la comprensión de la información por medio de algoritmos y también por el procesamiento del lenguaje natural, esto con el fin de dar una semiestructura a toda esa información por medio de un formato XML. El objetivo del segundo trabajo consiste en analizar una variedad en textos con diferentes formatos, tamaños, idiomas y contextos, esto con la finalidad de manejar el análisis de la información por medio de patrones contextuales y relacionales.

La mayoría de los trabajos relacionados presentados abordan el problema de Big Data en general de acuerdo a sus 3 características principales las cuales son el volumen, la velocidad y la variedad, como por ejemplo el trabajo de (D. Rustem, 2017) y (N. Kaur, 2019), ambos están enfocados a modelar y caracterizar las 'tres V' de Big Data para proporcionar información sobre el procesamiento, gestión y análisis de los datos.

Tabla 1. Comparación de trabajos relacionados

Acontinuación se muestran algunos acrónimos como referencia para la Tabla 1:

- SMS: Systematic Mapping Study (Estudio de Mapeo Sistemático).
- HTA: Herramienta.
- FWK: Framework.
- MTG: Metodología.
- VOL: Volumen.
- VEL: Velocidad.
- VAR: Variedad.
- N/A: No Aplica

N°	Trabajo Relacionado	Aportación					Procesamiento/Estrategia	Enfoque		
		SMS	HTA	FWK	MTG	OTRO		VOL	VEL	VAR
1	Systematic Mapping Studies in Software Engineering	**X**					Presenta un procesamiento para realizar un mapeo sistemático por medio de 5 pasos: 1. Definición de la pregunta de investigación. 2. Conducción de la búsqueda. 3. Selección de estudios. 4. Esquema de clasificación. 5. Extracción y síntesis de datos.	**N/A**	**N/A**	**N/A**
2	Descubrimiento de Conocimiento en Big Data: Estudio de Mapeo Sistémico	**X**					El proceso utilizado para realizar el mapeo sistemático fue por (K. Petersen F. R., 2008): 1. Definición de las preguntas de investigación. 2. Selección de fuentes. 3. Conducción de la búsqueda. 4. Selección de estudios. 5. Extracción y síntesis de datos.	**N/A**	**N/A**	**N/A**
3	A Systematic Mapping Study in Microservice Architecture	**X**					El proceso utilizado para realizar el mapeo sistemático fue: 1. Definición de las Preguntas de investigación. 2. Estrategia de búsqueda. 3. Selección de estudios primarios. 4. Palabras clave y clasificación. 5. Criterios de Inclusión y exclusión.	**N/A**	**N/A**	**N/A**

							6. Estrategia de extracción de datos y evaluación de la calidad.			
4	Research on Big Data - A systematic mapping study	**X**					El proceso utilizado para realizar el mapeo sistemático fue: 1. Definición de las preguntas de investigación. 2. Alcance de la revisión. 3. Búsqueda de estudios. 4. Selección de estudios. 5. Proceso de Inclusión y exclusión de estudios. 6. Revisión del Abstract por palabras clave. 7. Esquema de clasificación. 8. Extracción y síntesis de datos. 9. Mapeo Sistemático.	**N/A**	**N/A**	**N/A**
5	A Generic Framework for Concept-Based Exploration of Semi-Structured Software Engineering Data	**X**					El proceso utilizado para realizar el mapeo sistemático fue: 1. Definición de las Preguntas de investigación. 2. Selección de estudios. 3. Enfoque técnico. 4. Construcción del concepto formal 5. Implementación de prototipos 6. Escalabilidad 7. Visualizaciones de datos			**X**
6	Tactical Big Data Analytics: Challenges, Use Cases, and Solutions			**X**			Enfoque sistemático basado en la computación en la nube para proporcionar algoritmos de análisis y minería de datos escalables, así como herramientas y plataformas para analizar datos de sensores en tiempo real. Principales herramientas para el análisis de Big Data: - Hadoop Core - MapReduce - Apache Pig - HBase	**X**		**X**

							- Hive - Storm			
7	Variety Management for Big Data					**Ontología**	Introduce el concepto de metadatos semánticos como base para describir y gestionar la variedad en el contexto de Big Data desde 5 perspectivas: 1. Variación estructural. 2. Variaciones en la granularidad. 3. Fuentes de datos heterogéneas. 4. Grados de calidad e integridad. 5. Diferencias en los datos.			**X**
8	From Text to XML by Structural Information Extraction					**Algoritmo**	Plantea un algoritmo de extracción de datos por medio de la estructura del texto libre y su conversión al formato XML para su análisis y extracción de conocimiento.			**X**
9	Adaptive System for Handling Variety in Big Text					**Sistema / Modelo**	El sistema propuesto maneja una variedad de texto en un solo modelo. El sistema gestiona el texto de diversas fuentes, formatos, lenguajes y estructuras dentro del ecosistema de Big Data por medio de clasificadores y construcciones de modelos de texto.			**X**
10	A general perspective of Big Data: applications, tools, challenges and trends		**X**				Determinan una clasificación de herramientas para Big Data basadas en el análisis por lotes y análisis interactivo como son: -Google MapReduce -Apache Hadoop, Mahout, Storm, S4, Spark -Microsoft Dryad -Stream analysis -MOA -Apache Drill, SpagoBI y D3.js	**X**	**X**	**X**

11	Big Data Validation Case Study		**X**				Para el proceso de validación de Big Data realizó: -Recopilación, Limpieza, Transformación, Carga, Validación y Análisis de datos. El caso de estudio se llevó a cabo de la siguiente manera: -Descripción del caso de estudio -Herramientas de validación de Big Data -Resultados del caso de estudio Las herramientas empleadas para el análisis de Big Data fueron: -Datameer, Pentaho, ETL Processor, Talend y Querysurge		**X**	
12	Big Data: Framework and issues		**X**				Muestra como el análisis de Big Data por medio de herramientas ayudaría a las organizaciones empresariales a aumentar su valor por medio del análisis de la información, algunas de las herramientas descritas son: - Hadoop, HDFS, MapReduce, HBase, Pig, Hive, Mahout, Oozie, Sqoop, ZooKeeper, Ambari. -Apache Spark: Spark Core (API), Spark clustering y Spark stack	**X**	**X**	**X**
13	Big Data: Issues, Challenges, and Techniques in Business Intelligence		**X**				Sugieren dos métodos para el procesamiento de Big Data: - Utilizar Hadoop para el procesamiento de datos. - Utilizar Apache Spark para administrar los requisitos de procesamiento de datos de Big Data.	**X**	**X**	
14	Efforts toward Research and Development on Inconsistencies		**X**				El uso de herramientas depende principalmente de su finalidad y necesidad de la empresa, las herramientas más utilizadas son: - Pentaho, TerraEchos, Cognos, Attivio - Google BigQuery, Netezza, Apache Hadoop	**X**	**X**	**X**

	and Analytical tools of Big Data						- Zettaset, HPCC Systems, Dremel - MarketGreenplum HD, HortonWorks - ParAccel y GridGrain			
15	Evolution of Spark Framework for simplifying Big Data Analytics			**X**		**Modelo**	Propone un modelo donde Spark se pueda integrar con un conjunto de herramientas para el procesamiento de datos por medio de capas, algunas herramientas propuestas son: - Hadoop, MLBase, GraphX, Spark Streaming - Apache Kafka y Spark SQL	**X**	**X**	
16	Big Data and the SP Theory of Intelligence					**Modelo**	Proporciona un manejo del problema de la veracidad, velocidad y volumen en Big Data para ayudar en la gestión de errores e incertidumbres en los datos. Técnicas utilizadas: - Procesamiento del lenguaje natural - Reconocimiento de patrones		**X**	
17	A Comparative Study to Classify Big Data Using Fuzzy Techniques		**X**			**Modelo**	Presentan un estudio comparativo de clasificación en Big Data, utilizando tres algoritmos de clasificación; el dato k más cercano, el dato difuso k más cercano y la máquina de vectores de soporte que utiliza MapReduce. Proponen modelos matemáticos y algoritmos para el procesamiento difuso de datos en Big Data.		**X**	
18	A Big Data-as-a-Service Framework: State-of-the-Art and Perspectives			**X**			Presentan un nuevo framework de Big Data como servicio para representar, reducir, integrar y procesar Big Data.	**X**	**X**	**X**
19	Quantifying Volume, Velocity, and Variety to			**X**			Proponen un framework para proporcionar una estimación de las métricas de los V-atributos mediante la evaluación de un modelo de	**X**	**X**	

	Support (Big) Data-Intensive Application Development						rendimiento generado a partir del proceso de datos en Big Data.			
20	BIG Data and Methodology-A review				**X**		Muestra las tecnologías para el análisis de los datos más populares: - Hadoop - MapReduce - HDFS - HBase - Pig - Hive - Sqoop - Avro - Oozie - Chukwa - Flume - Zookeepe	**X**		**X**
21	Cloud resource management using 3Vs of Internet of Big Data streams					**Método**	Método propuesto que mejora el rendimiento de transmisión de datos de IoBd a través de la nube en tiempo real. El método trabaja en dos módulos llamados: - Workcaster Forecaster (Wo-Fo). - Resource Manager (RM).	**X**	**X**	**X**
22	Big Data Reduction Methods: A Survey					**Método**	Métodos de reducción de Big Data: - Teoría de la red. - Compresión. - Eliminación de redundancia. - Pre procesamiento de datos. - Reducción de dimensiones. - Minería de datos y aprendizaje automático		**X**	
23	Metodología para el modelamiento de datos basado en Big Data, enfocados al consumo de tráfico (voz-datos)		**X**		**X**		Metodología propuesta para el análisis de información enfocado al consumo de tráfico (Voz-Datos) en Big Data. Herramientas utilizadas: - Hadoop: para el almacenamiento y procesamiento de información. - Tableau: para analizar comportamiento de los datos de una forma intuitiva.	**X**	**X**	

	generado por los clientes									
24	A Big Data methodology for categorising technical support requests using Hadoop and Mahout			**X**		**Modelo**	Propone una solución para realizar análisis a gran escala de datos de Big Data utilizando el framework Hadoop con HBase y Hive, también utiliza algoritmos de agrupamiento por medio de Mahout.		**X**	**X**

2.3 Resumen del capítulo

En este capítulo se mostraron un conjunto de estudios que están enfocados al análisis y procesamiento de datos de Big Data, algunos estudios tienen enfoque matemático y algorítmico para el análisis estructural de los datos con ayuda de diferentes técnicas de procesamiento de información y otros estudios se enfocan al análisis de datos por medio de herramientas y frameworks con el fin de desarrollar un modelo o sistema que ayude al procesamiento de datos. En el siguiente capítulo se presentan los términos necesarios para tener una mejor comprensión de este trabajo de tesis.

Capítulo 4

Metodología

En este capítulo se representan las actividades que se llevaron a cabo para realizar el estudio de mapeo sistemático, se definió un objetivo principal, el cual fue transformado en las preguntas de investigación, que nos guiaron a las siguientes etapas del estudio hasta llegar a un bosquejo de selección bien definido, con el cual se pudo abordar el tema del problema de la variedad en sistemas Big Data.

En el estudio realizado se desarrolló el mapeo por medio de cinco procesos deacuerdo a (K. Petersen F. R., 2008) los cuales se describen a continuación:
(1) Definición de las Preguntas de Investigación, (2) Selección de Fuentes, (3) Conducción de la Búsqueda, (4) Selección de Estudios, (5) Extracción y Síntesis de Datos.

Estos procesos pertenecen a unas fases conocidas como Planeación (Protocolo y Preguntas), Conducción (Fuentes, Búsqueda y Selección) y Resultados (Extracción y Síntesis).

La Figura 1 esquematiza el proceso del mapeo. Posteriormente se detalla cada una de las fases.

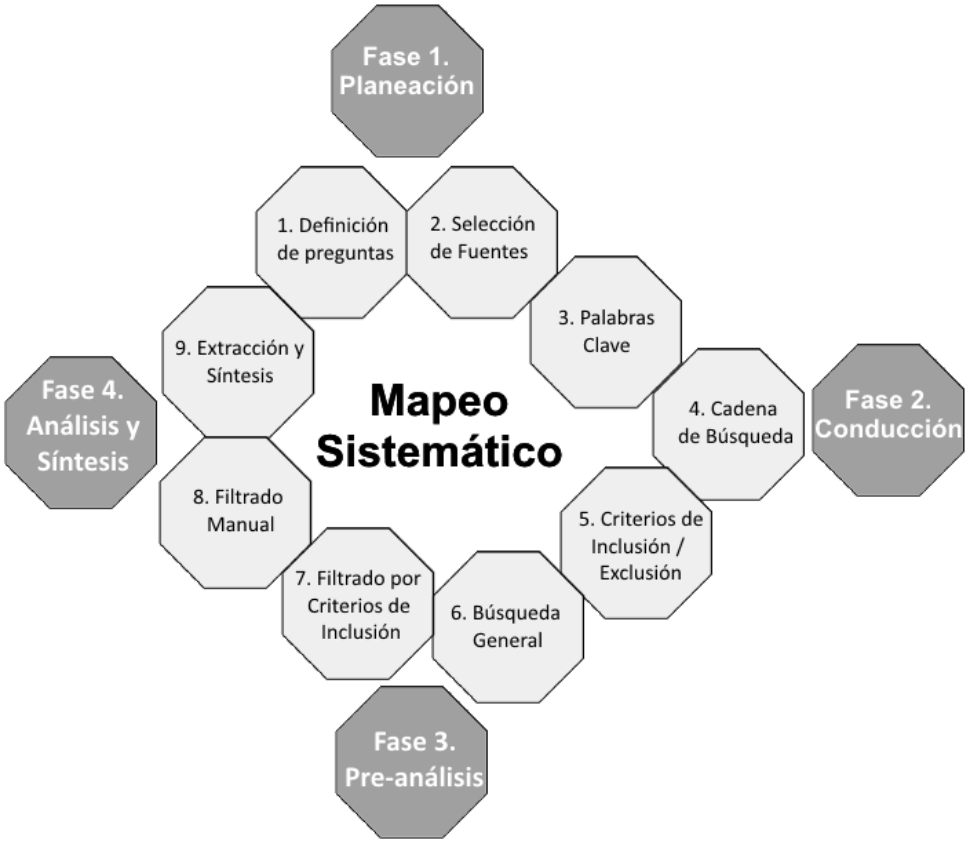


Figura 1. Proceso del mapeo sistemático adaptado (K. Petersen F. R., 2008)

4.1 Fase 1 – Planeación

En la primera fase se definió el problema a través del planteamiento de preguntas de investigación que orientaran las fases subsecuentes desde la búsqueda hasta el análisis de la información. Se procuró que las preguntas posibilitaran una navegación amplia dentro del tema de Big Data. Se plantearon tres preguntas de investigación.

4.1.1 Preguntas de Investigación

Las preguntas fundamentales del estudio fueron las siguientes:
RQ= ¿Existe algún estudio de mapeo sistemático que aborde el problema de la variedad? y ¿Qué herramientas, framework y metodologías son empleados para abordar el problema de la variedad en sistemas Big Data?, de la cual se derivó en 4 preguntas de investigación que se muestran en la Tabla 2.

Tabla 2. Preguntas de Investigación

ID	Pregunta	Explicación
RQ1	¿Existe algún estudio de mapeo sistemático que aborde el problema de la variedad en Big Data?	Busca evidenciar los estudios de mapeos sistemáticos para abordar el problema de la variedad en Big Data.
RQ2	¿Qué herramientas son empleadas para abordar el problema de la variedad en Big Data?	Busca evidenciar las herramientas que son empleadas para abordar el problema de la variedad en Big Data.
RQ3	¿Qué framework son utilizados para abordar el problema de la variedad en sistemas Big Data?	Busca demostrar y dar a conocer la existencia de algún framework que ayude a resolver el problema de la variedad en Big Data.
RQ4	¿Qué Metodologías son empleadas para abordar el problema de la variedad en sistemas Big Data?	Busca demostrar la distribución de los artículos seleccionados y analizarlos para saber si existe alguna metodología empleada o desarrollada que ayude a resolver el problema de la variedad en Big

		Data.

Para la formulación de la pregunta de investigación se revisó y aplicó la técnica PICO (K. Barbara Kitchenham, 2007), de la siguiente forma:

- **Population**: Todos los estudios publicados que aborden el problema de la variedad de Big Data con ayuda de Herramientas, frameworks y Metodologías.
- **Intervention:** La intervención de la búsqueda de las herramientas, frameworks y herramientas es estableciendo las palabras clave que guíen al estudio, esto de acuerdo al objetivo mencionado.
- **Control:** Criterios de inclusión y exclusión de artículos.
- **Outcome:** Listado de evidencias concretas sobre Herramientas, frameworks y Metodologías obtenidas para abordar el problema de la variedad en Sistemas Big Data.

4.1.2 Selección de Fuentes

Las fuentes utilizadas en este estudio fueron las siguientes bibliotecas digitales: ACM digital library, IEEE Xplore, ScienceDirect, JSTOR y SpringerLink. Los criterios de selección utilizados en este estudio fueron básicamente la disponibilidad de estas bibliotecas y el hecho de ser referentes en Ciencias de la Computación. Con respecto a la disponibilidad, el Centro Nacional de Investigación y Desarrollo Tecnológico (Cenidet) tiene un convenio con Consorcio Nacional de Recursos de Información Científica y Tecnológica (CONRICYT) para el acceso de los estudiantes. La Tabla 3 muestra la información de las fuentes seleccionadas.

Tabla 3. Información de las fuentes seleccionadas

Fuente	Dirección URL
Acm Digital Library	https://dl.acm.org/
ScienceDirect	https://www.sciencedirect.com/
IEEE Xplore	https://ieeexplore.ieee.org
JSTOR	https://www.jstor.org/
SpringerLink	https://link.springer.com/

4.2 Fase 2 – Conducción

En la segunda fase se definió la estrategia de búsqueda donde se establecen las palabras clave y sus sinónimos, una vez establecidas las palabras claves se generan y organiza la cadena de búsqueda y por último se realiza una búsqueda general de acuerdo a las fuentes seleccionadas previamente.

4.2.1 Palabras Clave

A partir del objetivo y las preguntas de investigación se obtuvieron las siguientes palabras clave: **Variedad, Herramienta, Framework, Metodología y Big Data**.

4.2.2 Sinónimos

Se buscaron y generaron sinónimos para las palabras clave en idioma inglés. En este caso, se acudió a los estudios analizados previamente sobre los tópicos para identificar los términos con los que son referidas normalmente las palabras clave, los cuales se muestran en la Tabla 4.

Tabla 4. Palabras Clave y sus sinónimos / combinación

Palabra Clave	**Sinónimos**
Tools	IDE
Framework	Frame
	Alternativa
Methodology	Method
	Combinación de palabras
Big Data	Big Data Systems
Variety	Variety Problem

4.2.3 Cadena de Búsqueda

Se organizó la cadena de búsqueda “SQ” con las palabras clave utilizando los operadores OR para sinónimos o alternativas, AND para combinar las palabras clave y NOT para exclusiones o negaciones.

La primera cadena de búsqueda "**SQ**" fue la base para buscar los artículos relacionados con el mapeo sistemático en el problema de variedad de Big Data.
SQ= ((Title:(A systematic AND (mapping OR map) AND (in variety problem OR in variety) AND ((Methodologies OR Method) OR (Framewoks OR Frame) OR (Tool OR IDE)) AND (Big Data OR Big Data systems))) OR (Abstract: (Systematic Mapping OR Map OR Variety Problem OR Variety in Big Data)) OR (Keywords: (Systematic Mapping OR Variety OR Variety problem OR Big Data Systems OR Big Data))).

Después de un refinamiento para la búsqueda, la cadena de búsqueda "SQ" se dividió en 4 cadenas específicas para la búsqueda de estudios.
La cadena "**SQ1**" fue utilizada para buscar los artículos relacionados con el mapeo sistemático específicamente con el problema de la variedad en Big Data, esto con el fin de saber si hay un estudio similar o previo a esta investigación.

SQ1= ((Title:(A systematic AND (mapping OR map) AND (in variety problem OR in variety) AND (Big Data OR Big Data systems))) OR (Abstract:(Systematic Mapping OR problem of variety OR variety in Big Data)) OR (Keywords:(Systematic Mapping OR Variety OR Variety problem OR Big Data systems OR Big Data)))

Para la búsqueda de herramientas fue un procedimiento similar, donde la primera cadena de búsqueda "**SQ1**" sólo hace mención del término de herramientas sin utilizar los sinónimos de las palabras clave. Para ello la cadena de búsqueda "**SQ2**" utiliza los sinónimos para la búsqueda de herramientas utilizadas en el problema de la variedad en Big Data.
SQ2= ((Tools OR IDE) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))

De igual manera las cadenas de búsqueda "SQ3" y "SQ4" utilizan los sinónimos y combinación de palabras para la búsqueda de frameworks y metodologías que aborden el problema de la variedad.
SQ3= ((Framework OR Frame) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))
SQ4= ((Methodology OR Method) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))

4.2.4 Criterios de Inclusión y Exclusión

Para la selección de los estudios, se definieron los criterios de inclusión y exclusión que se muestran en la Tabla 5, con los cuales se descartaron los estudios no relevantes o que no respondieron a alguna de las preguntas de investigación.

Tabla 5. Criterios de Inclusión / Exclusión

Tipo	Criterio
Inclusión	• Artículos publicados entre los años 2013 y 2019 • Artículos publicados en Journals y Conference Proceedings • Artículos escritos en idioma inglés • Artículos que tengan acceso a la revisión de abstract y keywords mínimo • Artículos que aborden el problema de variedad en Big Data • Artículos que permitan evidenciar el problema de variedad • Artículos que aborden el problema por medio de herramientas, frameworks o metodologías.
Exclusión	• Artículos completos. • Documentos que estén en un formato de presentaciones o textos informativos. • Trabajos de patentes o puramente teóricos sin implementación o comprobación. • Artículos que presenten resultados o conclusiones meramente cualitativas. • Estudios que no aborden el contexto de Big Data

4.3 Fase 3 – Pre-análisis

El pre-análisis conllevó el primer acercamiento a los documentos. Se leyeron los títulos y los resúmenes de cada texto, valorando la pertinencia de cada documento identificado en la búsqueda y seleccionado en el refinamiento. Se establecieron los criterios de inclusión y posteriormente los criterios de exclusión en la fase 4.

4.3.1 Generación de la Búsqueda en las fuentes seleccionadas (Búsqueda General).

Las cadenas de búsqueda generadas fueron ajustadas según el formato de cada fuente para su posterior aplicación en las mismas. Cabe resaltar que en este punto se aplicaron los siguientes criterios de inclusión: Años [2013, 2019]; Tipos de publicación {Papers, Journals, Conference Proceedings}; Idioma {inglés}; Acceso a Abstract y Keywords {Sí}. En la Tabla 6 se muestran las búsquedas ajustadas y realizadas en cada fuente, las cuales arrojaron 79,599 inicialmente para herramientas, para frameworks el resultado inicial fue de 2,278 estudios y 4,086 estudios para metodologías, dando un total de 85,963 estudios.

En el caso de la publicación de un estudio de mapeo sistemático con el problema de la variedad en Big Data el resultado fue de 0 estudios.

4.3.2 Filtrado por criterios de inclusión.

Se realizó el proceso de filtrado semiautomático para cada base de datos tomando en cuenta los criterios de inclusión y exclusión. Con este filtrado se llegó a 5,175 estudios para herramientas equivalente a 6.50% de los estudios iniciales, 540 estudios para frameworks equivalente a 23.7% y 415 estudios para metodologías equivalente a 10.15%, dando un total de 6,130 en su totalidad representa un 7.13% del total de los estudios generales arrojados inicialmente. En la Tabla 7 se muestran el resultado del filtrado semiautomático de duplicados.

4.4 Fase 4 – Análisis y Síntesis

La fase 4 conllevó el filtrado manual de los estudios. Se leyeron el total de artículos para su posterior clasificación, valorando el impacto, el uso de herramientas, frameworks, metodologías, algoritmos y métodos propuestos para abordar la variedad en el contexto de Big Data. Se establecieron los criterios de exclusión y posteriormente una clasificación de estudios relevantes.

4.4.1 Filtrado por criterios de exclusión.

Utilizando los mismos criterios del punto anterior (4.3.2), se descartaron los estudios que no tenían relevancia al no abordar los objetivos de las preguntas de investigación quedando un 1.43% de estudios relevantes. Con este filtrado se llegó a 88 estudios, los cuales son 37 para herramientas, 24 para frameworks y 27 para metodologías. En la Tabla 8 se muestran el resultado del filtrado manual de estudios relevantes.

4.4.2 Resultado de las cadenas de búsqueda por fuente seleccionada.

4.4.2.1 Resultado de la cadena y búsqueda sin ajustes.

Tabla 6. Cadena y Búsqueda de estudios sin ajustes.

Fuente	Cadena de Búsqueda SQ1 “Mapeo Sistemático”	Cantidad de estudios
ACM DL	+((Title:("A systematic" +("mapping" "map") +(“in variety problem” “in variety”) +(“Big Data” “Big Data systems”))) +(Abstract:("Systematic Mapping" "problem of variety" "variety in Big Data")) +(Keywords:("Systematic Mapping" “Variety” “Variety problem” "Big Data systems" "Big Data")))	0
ScienceDirect	((Title:(A systematic AND (mapping OR map) AND (in variety problem OR in variety) AND (Big Data OR Big Data systems))) OR (Abstract:(Systematic Mapping OR problem of variety OR variety in Big Data)) OR (Keywords:(Systematic Mapping OR Variety OR Variety problem OR Big Data systems OR Big Data)))	0
IEEE Xplore		0
JSTOR		0
SpringerLink		0

<table>
<tr><td colspan="3">• Total SQ1: 0 estudios</td></tr>
<tr><td>Fuente</td><td>Cadena de Búsqueda <u>SQ2 “Herramientas”</u></td><td>Cantidad de estudios</td></tr>
<tr><td>ACM DL</td><td>+(“Tools” “IDE”) +(“variety” “variety problem”) +(“Big Data” “Big Data Systems”))</td><td>262</td></tr>
<tr><td>ScienceDirect</td><td rowspan="4">((Tools OR IDE) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))</td><td>71,069</td></tr>
<tr><td>IEEE Xplore</td><td>354</td></tr>
<tr><td>JSTOR</td><td>1,384</td></tr>
<tr><td>SpringerLink</td><td>6,530</td></tr>
<tr><td colspan="3">• Total SQ2: 79,599 estudios</td></tr>
<tr><td>Fuente</td><td>Cadena de Búsqueda <u>SQ3 “Frameworks”</u></td><td>Cantidad de estudios</td></tr>
<tr><td>ACM DL</td><td>+(“Framework” “Frame”) +(“variety” OR “variety problem”) +(“Big Data” “Big Data Systems”))</td><td>209</td></tr>
<tr><td>ScienceDirect</td><td rowspan="4">((Framework OR Frame) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))</td><td>686</td></tr>
<tr><td>IEEE Xplore</td><td>178</td></tr>
<tr><td>JSTOR</td><td>157</td></tr>
<tr><td>SpringerLink</td><td>1,048</td></tr>
<tr><td colspan="3">• Total SQ3: 2,278 estudios</td></tr>
<tr><td>Fuente</td><td>Cadena de Búsqueda <u>SQ4 “Metodologías”</u></td><td>Cantidad de estudios</td></tr>
<tr><td>ACM DL</td><td>+(“Methodology” “Method”) +(“variety” “variety problem”) +(“Big Data” “Big Data Systems”)</td><td>128</td></tr>
<tr><td>ScienceDirect</td><td rowspan="4">((Methodology OR Method) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))</td><td>912</td></tr>
<tr><td>IEEE Xplore</td><td>80</td></tr>
<tr><td>JSTOR</td><td>86</td></tr>
<tr><td>SpringerLink</td><td>2,880</td></tr>
<tr><td colspan="3">• Total SQ4: 4,086 estudios</td></tr>
<tr><td colspan="3">• Total: 85,963 estudios iniciales</td></tr>
</table>

4.4.2.2 Resultado de la cadena y búsqueda por criterios de inclusión.

Tabla 7. Filtrado por criterios de inclusión.

Fuente	Cadena de Búsqueda SQ2 "Herramientas"	Cantidad de estudios
ACM DL	+("Tools" "IDE") +("variety" "variety problem") +("Big Data" "Big Data Systems"))	189
ScienceDirect	((Tools OR IDE) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))	4,015
IEEE Xplore		278
JSTOR		72
SpringerLink		621
		• Total SQ2: 5,175 estudios
Fuente	**Cadena de Búsqueda SQ3 "Frameworks"**	**Cantidad de estudios**
ACM DL	+("Framework" "Frame") +("variety" OR "variety problem") +("Big Data" "Big Data Systems"))	62
ScienceDirect	((Framework OR Frame) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))	200
IEEE Xplore		149
JSTOR		20
SpringerLink		109
		• Total SQ3: 540 estudios
Fuente	**Cadena de Búsqueda SQ4 "Metodologías"**	**Cantidad de estudios**
ACM DL	+("Methodology" "Method") +("variety" "variety problem") +("Big Data" "Big Data Systems")	107
ScienceDirect	((Methodology OR Method) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))	116
IEEE Xplore		29
JSTOR		34
SpringerLink		129
		• Total SQ4: 415 estudios
	• Total: 6,130 estudios por criterios de inclusión	

4.4.2.3 Resultado de la cadena y búsqueda por criterios de exclusión.

Tabla 8. Filtrado por criterios de exclusión.

Fuente	Cadena de Búsqueda SQ2 "Herramientas"	Cantidad de estudios
ACM DL	+("Tools" "IDE") +("variety" "variety problem") +("Big Data" "Big Data Systems"))	10
ScienceDirect	((Tools OR IDE) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))	2
IEEE Xplore		9
JSTOR		7
SpringerLink		9
		• Total SQ2: 37 estudios
Fuente	**Cadena de Búsqueda SQ3 "Frameworks"**	**Cantidad de estudios**
ACM DL	+("Framework" "Frame") +("variety" OR "variety problem") +("Big Data" "Big Data Systems"))	8
ScienceDirect	((Framework OR Frame) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))	5
IEEE Xplore		6
JSTOR		1
SpringerLink		4
		• Total SQ3: 24 estudios
Fuente	**Cadena de Búsqueda SQ4 "Metodologías"**	**Cantidad de estudios**
ACM DL	+("Methodology" "Method") +("variety" "variety problem") +("Big Data" "Big Data Systems")	10
ScienceDirect	((Methodology OR Method) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))	2
IEEE Xplore		5
JSTOR		0
SpringerLink		10
		• Total SQ4: 27 estudios
• Total: 88 estudios relevantes		

La Figura 2 muestra el resultado obtenido de la cantidad de estudios de acuerdo a las cadenas de búsqueda sin ajustes, por criterios de inclusión y por criterios de exclusión.

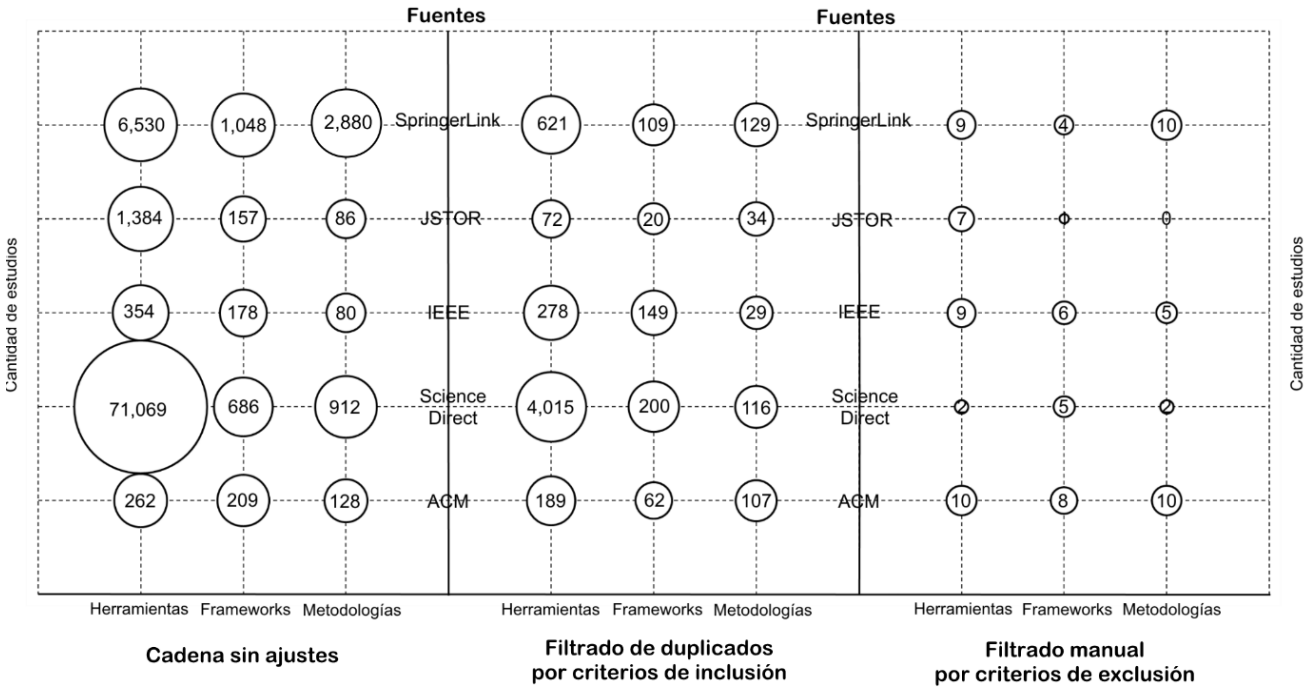


Figura 2. Diagrama de Burbujas del resultado de artículos por fuente

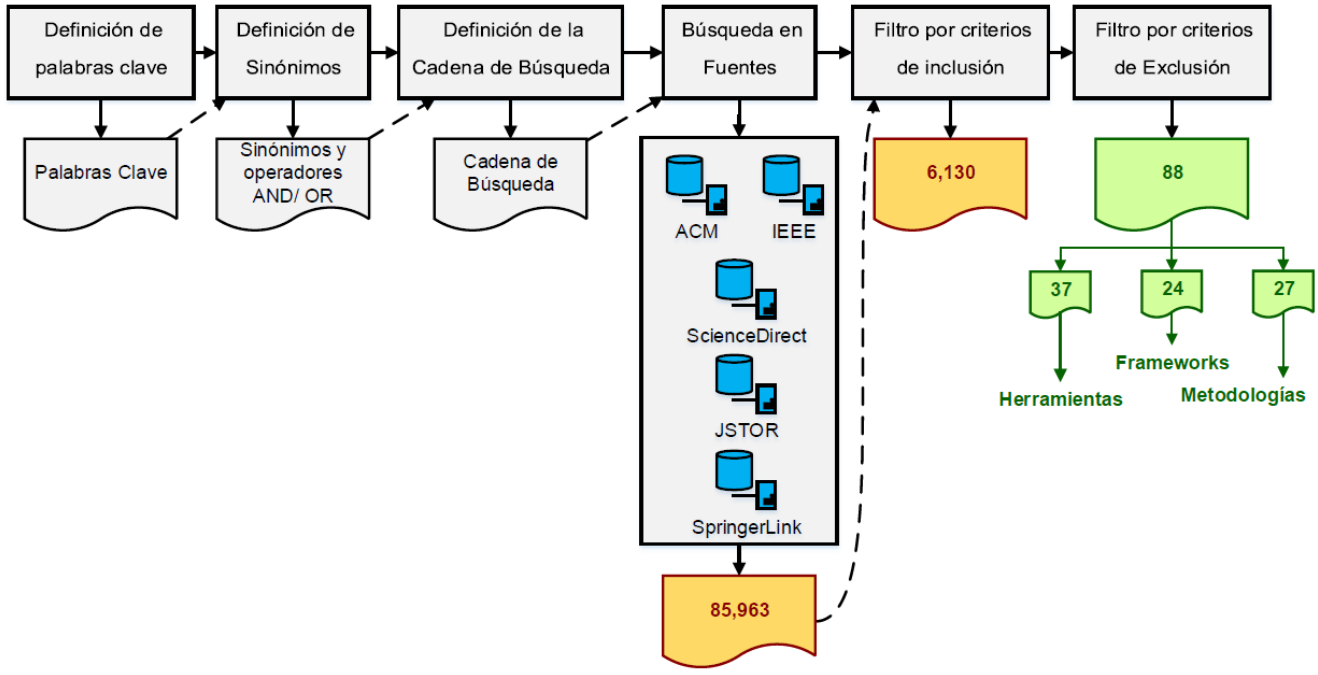


Figura 3. Proceso llevado acabo para la búsqueda de la información.

La Figura 3 muestra una vista general del proceso de mapeo sistemático donde se observan los cinco subprocesos:

(1) Definición de las Preguntas de Investigación: SQ, SQ1, SQ2, SQ3 y SQ4
(2) Selección de Fuentes: ACM, IEEE explore, ScienceDirect, JSTOR y SpringerLink
(3) Conducción de la Búsqueda: obtención de los primeros artículos 85,963
(4) Selección de Estudios: obtención de los artículos por medio de criterios de inclusión 6,130 y por criterios de exclusión 88.
(5) Extracción y Síntesis de Datos: 37 herramientas, 24 frameworks y 27 metodologías, desarrollo de la ampliación del estudio del arte y resultados obtenidos para abordar el problema de la variedad de acuerdo a la información obtenida.

4.5 Categoría de clasificación

La clasificación de los estudios se realizó en cuatro fases. En la primera fase, se aplicó la cadena de búsqueda general sin ningún tipo de filtro, esto con el fin de tener un panorama general sobre los estudios que abordan el problema de Big Data de acuerdo con algún mapa, mapeo o alguna herramienta, framework o metodología.

En la segunda fase se realizó un refinamiento de la búsqueda en las fuentes de datos establecidas y se agregó el filtrado por medio de las palabras clave dentro de la cadena de búsqueda, esto con el fin de delimitar la búsqueda de estudios. También se discriminaron los estudios que no tuvieron un contexto respecto a la característica de la variedad en Big Data. También se leyeron los títulos, palabras clave y resúmenes de los estudios.

En la tercera fase se eligieron los estudios principales de acuerdo a los criterios de exclusión y área de conocimiento (problema de la variedad) en el contexto de Big Data, a su vez se estudiaron los artículos por: resumen, introducción y conclusiones, esto con el fin de descartar artículos que no aportaban alguna posible estrategia, solución o aportación al problema de la variedad. También se obtuvieron algunas palabras secundarias que ayudaron a la clasificación de los estudios: Tipo de dato (Estructurado, No estructurado, Semi estructurado, homogéneo), Procesamiento de datos (Lotes, Stream), Formato de datos, etc.

En la cuarta fase se eligieron un conjunto final de artículos, se analizaron los estudios y se clasificaron de acuerdo por:

(1) Tipo: Herramienta, Framework o Metodología.
(2) Tipo de dato: Estructurado, Semiestructurado y No estructurado.
(3) Tratamiento de dato: Formato de Entrada/Salida

La Figura 4 muestra una vista general del proceso de selección de los documentos:

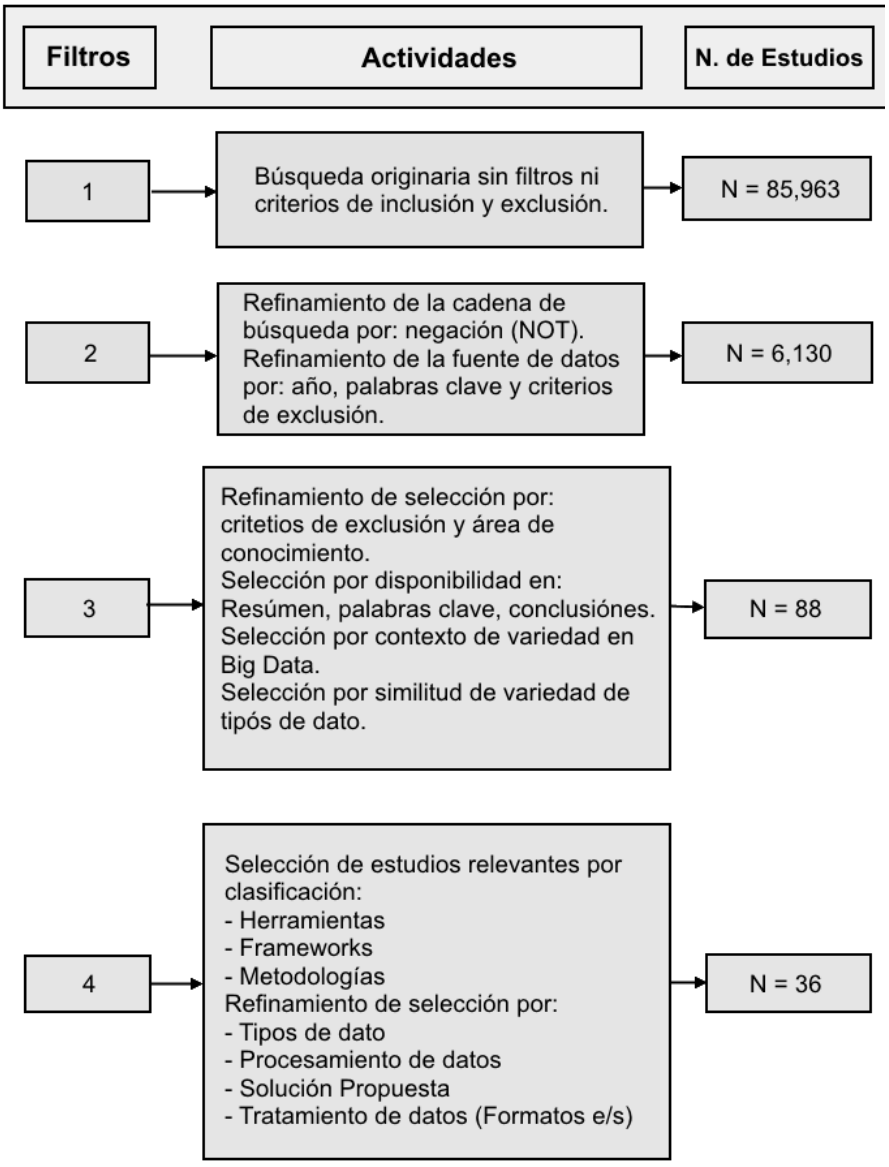


Figura 4. Proceso de selección de los documentos.

De acuerdo a (D. Gough, 2012), proponen un sistema de clasificación, centrándose en las variaciones en los objetivos, enfoques, estructuras, componentes, amplitud y profundidad de las revisiones de la investigación.

La Tabla 9 muestra un conjunto de categorías que ayudaron a organizar los artículos para realizar la revisión sistemática de los estudios. (D. Gough, 2012)

Tabla 9. Clasificación de enfoques de investigación

Categoría	Descripción
Investigación Exploratoria	Son estudios que suelen ser el primer acercamiento científico a un problema. Se utiliza cuando éste aún no ha sido abordado o no ha sido suficientemente estudiado y las condiciones existentes no son aún determinantes.
Investigación Explicativa	Con este tipo de investigación es posible encontrar la relación existente entre la causa y consecuencia de un problema. De esta forma es posible conocer el porqué de este y cómo ha llegado a su estado actual.
Investigación Aplicativa	La investigación trata de resolver un determinado problema o planteamiento específico, enfocándose en la búsqueda y consolidación del conocimiento para su aplicación con un grupo de herramientas computacionales existentes para un determinado fin.
Artículos Teóricos	Estos artículos tienen como objetivo la obtención de conocimiento sin importar su posterior adaptación del modelo al que sea aplicado.
Artículos de Revisión	Consiste en un grupo de trabajos elaborados a partir de artículos originales previamente publicados. A partir de un tema que el autor seleccionó para investigar, buscar, identificar, recopilar y revisar los trabajos más recientes.

De acuerdo a la Tabla 9 se realizó una clasificación de estudios por tipo de investigación, la cual sirvió para validar, clasificar y analizar los artículos finales como se muestra a continuación:

- La investigación exploratoria: tiene como fin obtener estudios iniciales que aporten una visión general al problema de la variedad en Big Data.

- La investigación explicativa: tiene como fin obtener estudios que evidencien las causas y efectos actuales del problema de la variedad, principalmente estudios con enfoques al análisis estructural de los datos.
- La investigación aplicativa: tiene como fin obtener estudios que aborden el problema de la variedad por medio de herramientas y frameworks y de esta manera evidenciar su procesamiento y tratamiento de los diferentes tipos de datos.

- Artículos Teóricos: tiene como fin obtener estudios que aborden el problema de la variedad de Big Data por medio de alguna metodología, modelos matemáticos, algoritmos, métodos, etc.

- Artículos de Revisión: tiene como fin obtener estudios que aborden el problema de la variedad por medio de mapas, mapeos, revisiones sistemáticas, etc.

4.6 Resumen del capítulo

En este capítulo se presentó el proceso del mapeo sistemático realizado para la búsqueda, selección y clasificación de los artículos de investigación que representan la evidencia sobre algún estudio de mapeo sistemático, asi como las herramientas, frameworks y metodologías para abordar el problema de la variedad en Big Data. En el siguiente capítulo se presenta el resultado de este estudio de mapeo sistemático.

Capítulo 5

Resultados del Estudio de Mapeo Sistemático

En este capítulo se presenta un análisis visual (mapeo sistemático) acerca del uso de herramientas, frameworks y metodologías en el problema de la variedad en Big Data. Mediante este resumen se presenta el panorama general acerca de la evidencia existente respecto al tratamiento de los diferentes tipos de datos. También en esta sección se muestra un análisis estadístico.

5.1 Análisis estadístico de los estudios

La cadena inicial "SQ1" referente a la búsqueda de un mapeo sistemático en el problema de la variedad en Sistemas Big Data, tuvo un resultado de 0 estudios.

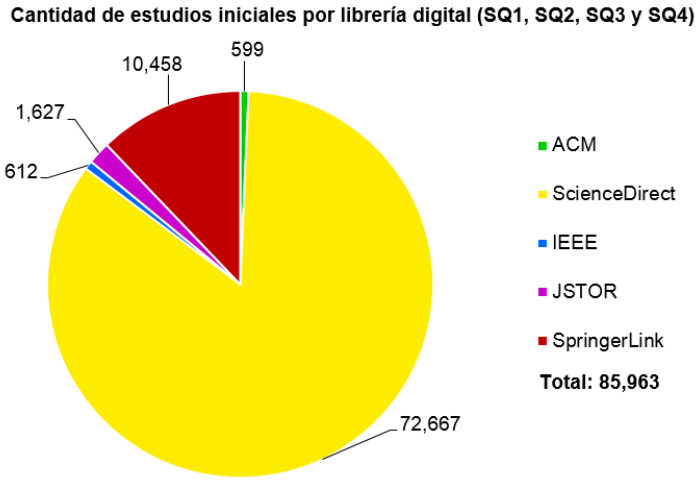
SQ1: ((Title:(A systematic AND (mapping OR map) AND (in variety problem OR in variety) AND (Big Data OR Big Data systems))) OR (Abstract:(Systematic Mapping OR problem of variety OR variety in Big Data)) OR (Keywords:(Systematic Mapping OR Variety OR Variety problem OR Big Data systems OR Big Data)))

Cabe mencionar que las cadenas de búsqueda fueron ajustadas según el formato de cada fuente y se realizó una búsqueda más refinada de cada cadena al incluir la búsqueda por título, resumen y palabras clave.

La Gráfica 1 muestra el total de artículos obtenidos con por las cadenas de búsqueda (SQ1, SQ2, SQ3 y SQ4) sin exclusiones con un total de 85,963 estudios iniciales, siendo ScienceDirect la fuente que arrojó mayor información con 72,667estudios y ACM la fuente que arrojó menor cantidad de información con 599 estudios.

Tabla 10. Total de estudios iniciales

Librería Digital	Artículos
ACM	599
ScienceDirect (ElSevier)	72,667
IEEE	612
JSTOR	1,627
SpringerLink	10,458
	∑= 85,963



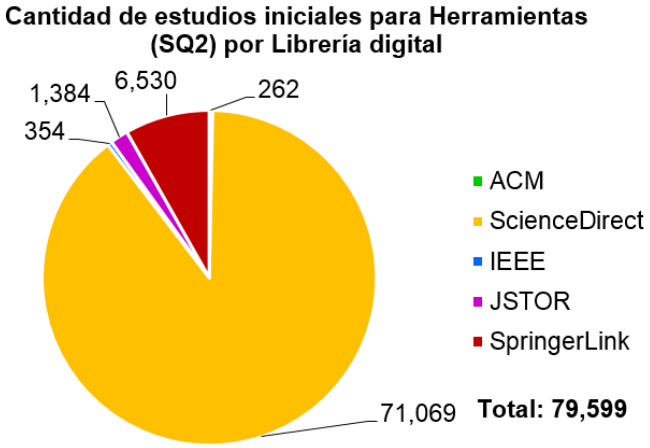
Gráfica 1. Total de estudios iniciales General

La Gráfica 2 muestra la cantidad de estudios obtenidos por la cadena de búsqueda SQ2, la cual obtuvo un total de 79,599 estudios iniciales para herramientas.

SQ2: ((Tools OR IDE) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))

Tabla 11. Total de estudios iniciales de herramientas

Librería Digital	Artículos
ACM	262
ScienceDirect (ElSevier)	71,069
IEEE	354
JSTOR	1,384
SpringerLink	6,530
	∑= 79,599



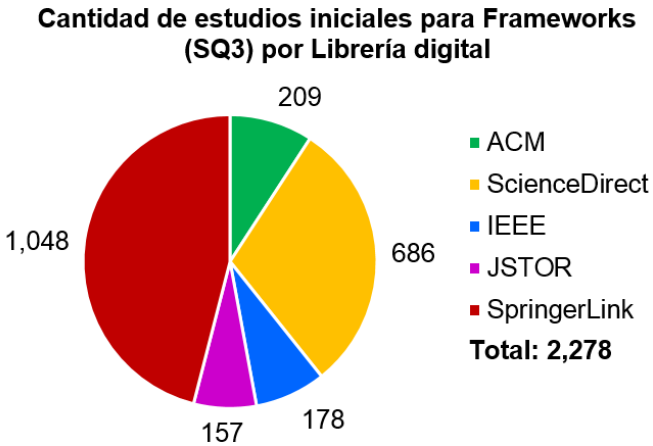
Gráfica 2. Total de estudios iniciales de herramientas

La Gráfica 3 muestra la cantidad de estudios obtenidos por la cadena de búsqueda SQ3, la cual obtuvo un total de 2,278 estudios iniciales para frameworks.

SQ3: ((Framework OR Frame) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))

Tabla 12. Total de estudios iniciales de frameworks

Librería Digital	Artículos
ACM	209
ScienceDirect (ElSevier)	686
IEEE	178
JSTOR	157
SpringerLink	1,048
	∑= 2,278



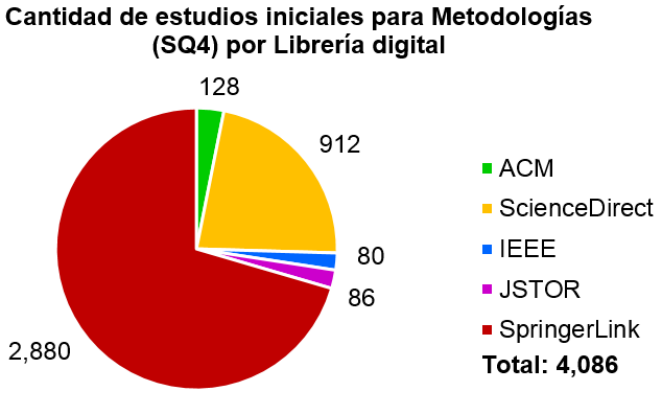
Gráfica 3. Total de estudios iniciales de frameworks

La Gráfica 4 muestra la cantidad de estudios obtenidos por la cadena de búsqueda SQ4, la cual obtuvo un total de 4,086 estudios iniciales para Metodologías.

SQ4: ((Methodology OR Method) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))

Tabla 13. Total de estudios iniciales de Metodologías

Librería Digital	Artículos
ACM	128
ScienceDirect (ElSevier)	912
IEEE	80
JSTOR	86
SpringerLink	2,880
	∑=4,086

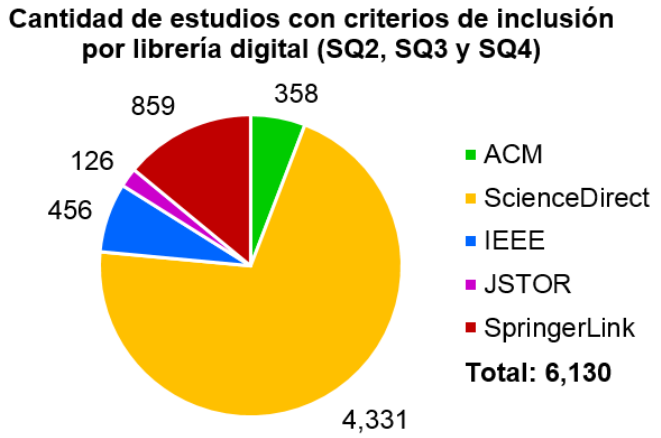


Gráfica 4. Total de estudios iniciales de Metodologías

La Gráfica 5 muestra el total de artículos obtenidos con por las cadenas de búsqueda (SQ2, SQ3 y SQ4) por los criterios de inclusión con un total de 6,130 estudios, siendo ScienceDirect la fuente que arrojó mayor información con 4,331 estudios y JSTOR la fuente que arrojó menor cantidad de información con 126 estudios.

Tabla 14. Total de estudios por criterios de inclusión

Librería Digital	Artículos
ACM	358
ScienceDirect (ElSevier)	4,331
IEEE	456
JSTOR	126
SpringerLink	859
	∑= 6,130

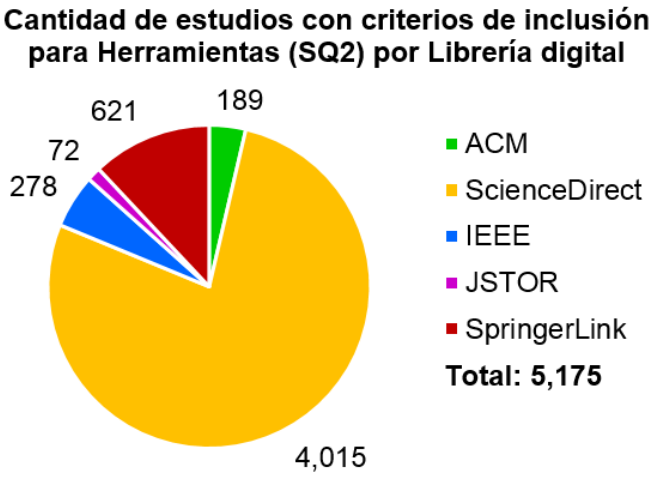


Gráfica 5. Total de estudios por criterios de inclusión

La Gráfica 6 muestra la cantidad de estudios obtenidos por la cadena de búsqueda SQ2 con los criterios de inclusión, la cual obtuvo un total de 5,175 estudios para herramientas.

Tabla 15. Total de herramientas por criterios de inclusión

Librería Digital	Artículos
ACM	189
ScienceDirect (ElSevier)	4,015
IEEE	278
JSTOR	72
SpringerLink	621
	∑= 5,175

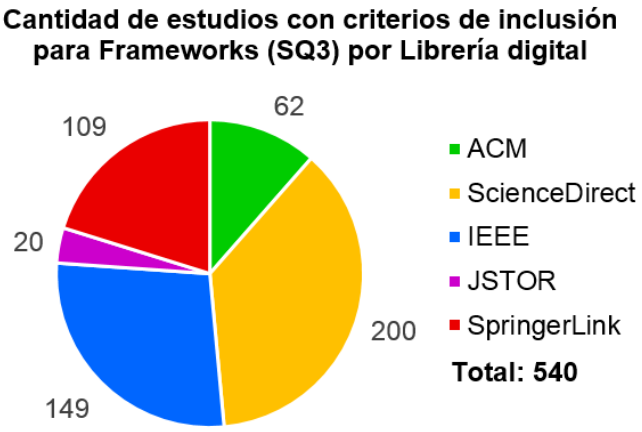


Gráfica 6. Total de herramientas por criterios de inclusión

La Gráfica 7 muestra la cantidad de estudios obtenidos por la cadena de búsqueda SQ3 con los criterios de inclusión, la cual obtuvo un total de 540 estudios para frameworks.

Tabla 16. Total de frameworks por criterios de inclusión

Librería Digital	Artículos
ACM	62
ScienceDirect (ElSevier)	200
IEEE	149
JSTOR	20
SpringerLink	109
	∑= 540

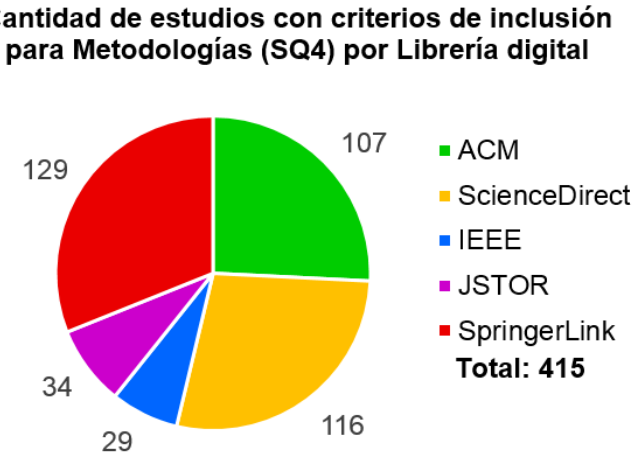


Gráfica 7. Total de frameworks por criterios de inclusión

La Gráfica 8 muestra la cantidad de estudios obtenidos por la cadena de búsqueda SQ4 con los criterios de inclusión, la cual obtuvo un total de 415 estudios para Metodologías.

Tabla 17. Total de Metodologías por criterios de inclusión

Librería Digital	Artículos
ACM	107
ScienceDirect (ElSevier)	116
IEEE	29
JSTOR	34
SpringerLink	129
	∑= 415

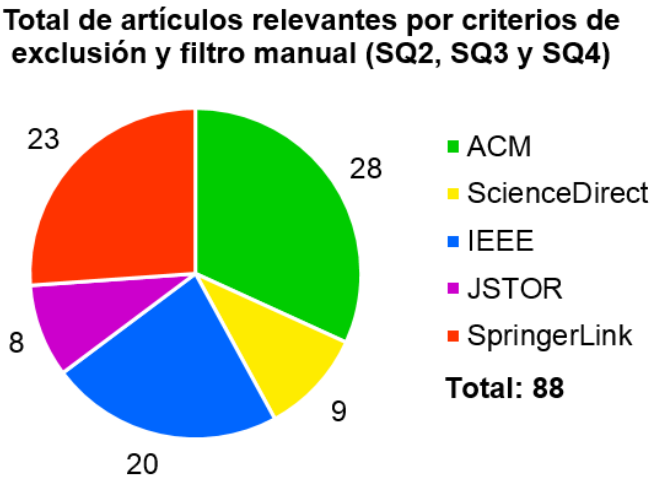


Gráfica 8. Total de metodologías por criterios de inclusión

La Gráfica 9 muestra el total de artículos obtenidos con por las cadenas de búsqueda (SQ2, SQ3 y SQ4) por medio de un filtro manual y por medio de los criterios de exclusión, dando un total de 88 estudios, de los cuales ACM la fuente la fuente con mayor información con 28 estudios y JSTOR la fuente con menor cantidad de información con 8 estudios.

Tabla 18. Total de artículos relevantes

Librería Digital	Artículos
ACM	28
ScienceDirect (ElSevier)	9
IEEE	20
JSTOR	8
SpringerLink	23
	∑= 88

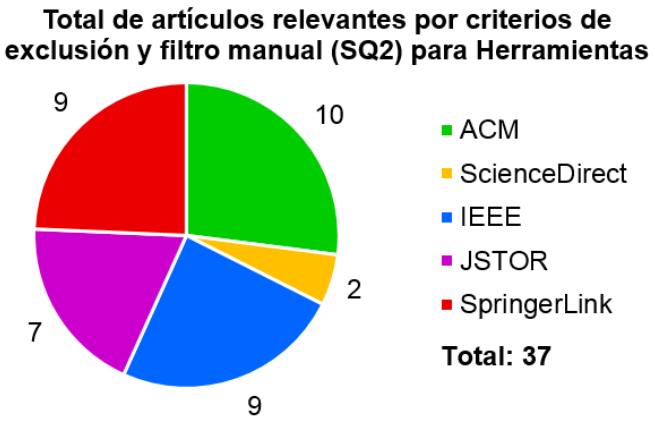


Gráfica 9. Total de artículos relevantes

La Gráfica 10 muestra la cantidad de estudios obtenidos por la cadena de búsqueda SQ2 por medio de un filtro manual y por medio de los criterios de exclusión, la cual obtuvo un total de 37 estudios para herramientas.

Tabla 19. Total de herramientas relevantes

Librería Digital	**Artículos**
ACM	10
ScienceDirect (ElSevier)	2
IEEE	9
JSTOR	7
SpringerLink	9
	∑= 37

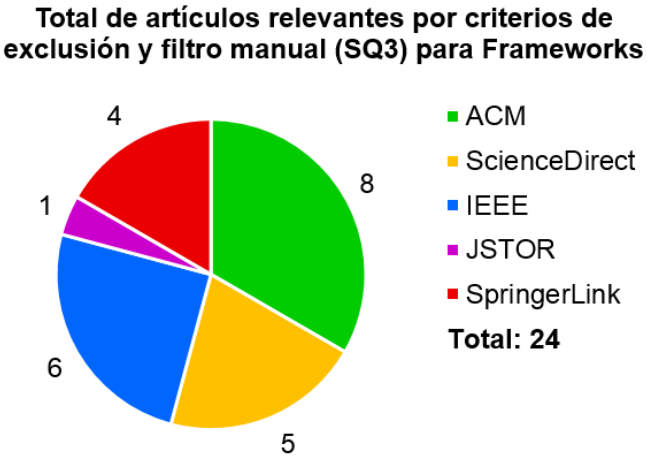


Gráfica 10. Total de herramientas relevantes

La Gráfica 11 muestra la cantidad de estudios obtenidos por la cadena de búsqueda SQ3 por medio de un filtro manual y por medio de los criterios de exclusión, la cual obtuvo un total de 24 estudios para frameworks.

Tabla 20. Total de frameworks relevantes

Librería Digital	**Artículos**
ACM	8
ScienceDirect (ElSevier)	5
IEEE	6
JSTOR	1
SpringerLink	4
	∑= 24

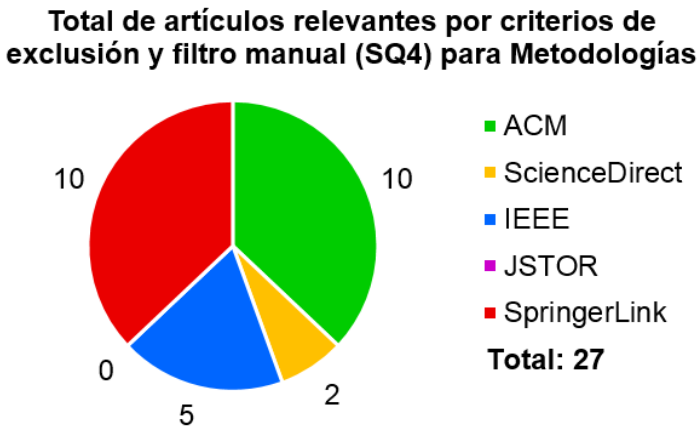


Gráfica 11. Total de frameworks relevantes

La Gráfica 12 muestra la cantidad de estudios obtenidos por la cadena de búsqueda SQ4 por medio de un filtro manual y por medio de los criterios de exclusión, la cual obtuvo un total de 24 estudios para Metodologías.

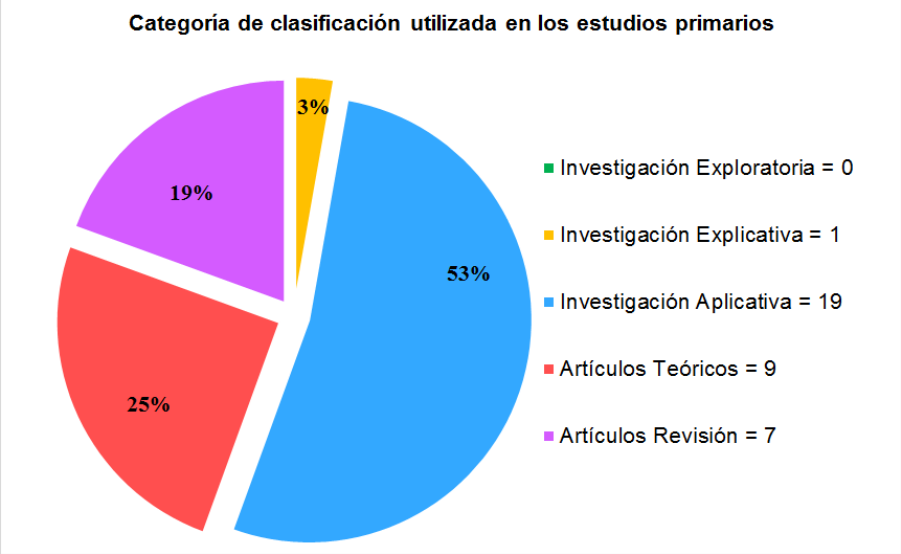
Tabla 21. Total de metodologías relevantes

Librería Digital	Artículos
ACM	10
ScienceDirect (ElSevier)	2
IEEE	5
JSTOR	0
SpringerLink	10
	∑= 27



Gráfica 12. Total de metodologías relevantes

La Gráfica 13 muestra la categoría de clasificación (ver categoría de clasificación) para cada estudio primarios. Se encontró que el 43% de los estudios fueron un conjunto de herramientas y frameworks para procesar diferentes tipos de datos, el 24% representan estudios que se basan en modelos teóricos como son algoritmos, modelos, métodos y metodologías, el 15% son estudios enfocados a un mapa, mapeo o revisión sistemática, el 11% son estudios que aportaron un análisis estructural para los diferentes tipos de datos, mientras que un 7% es representado por estudios que apotan una visión general al problema de la variedad.

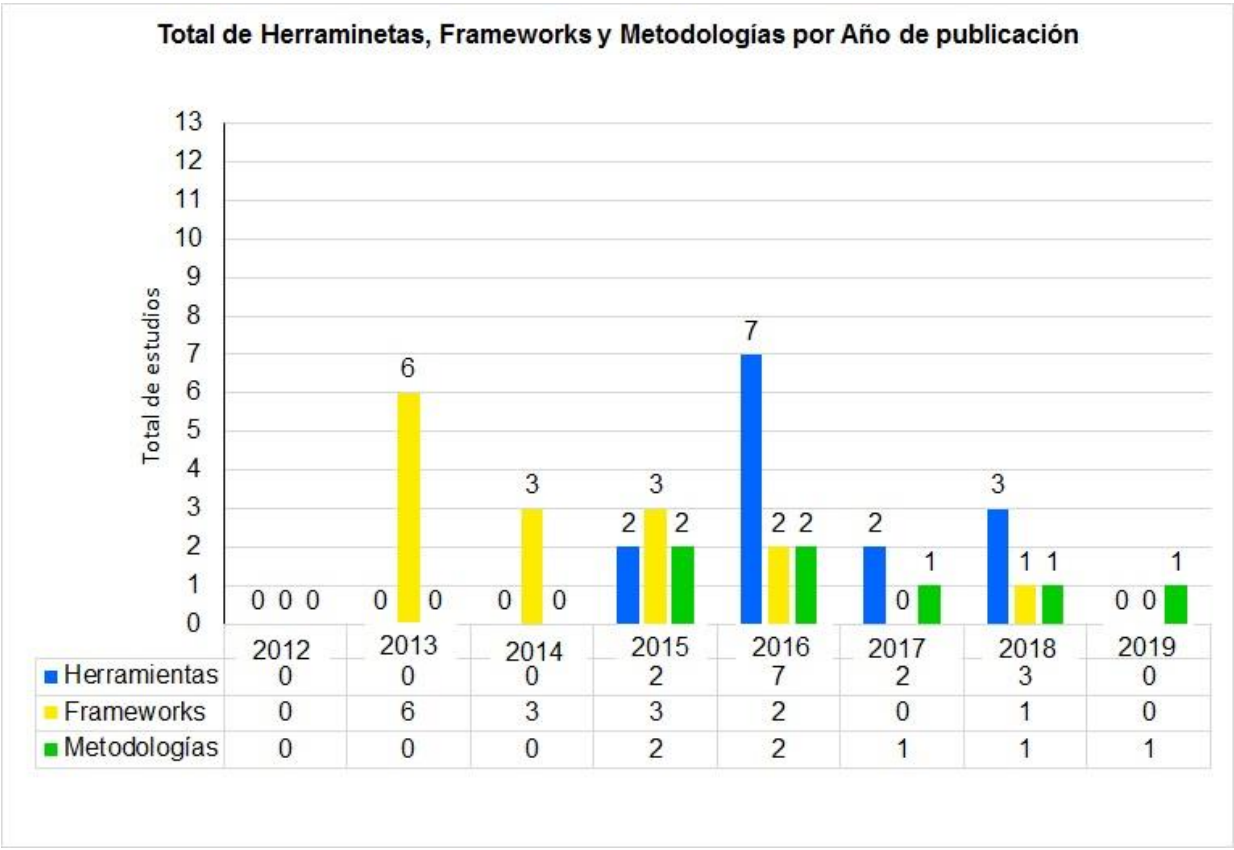


Gráfica 13. Total de estudios primarios por categoría de clasificación

La Gráfica 14 muestra la cantidad de estudios relevantes publicados por año, teniendo la mayor cantidad de artículos publicados en el año 2016 con un total de 28 de estudios.

Gráfica 14. Total de estudios relevantes por año de publicación

La Gráfica 15 muestra la cantidad de herramientas, frameworks y metodologías por año de publicación.



Gráfica 15. Cantidad de herramientas, frameworks y metodologías por año de publicación

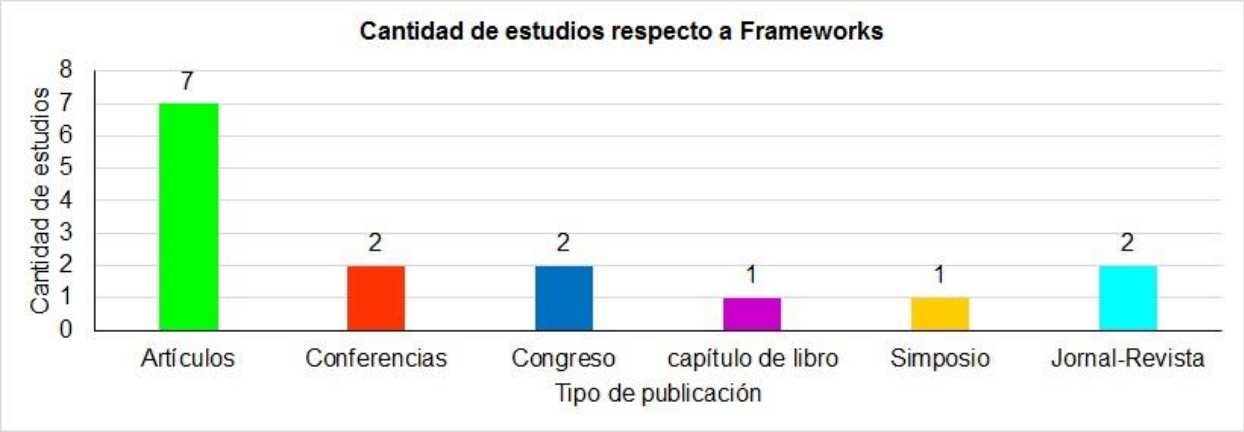
La Gráfica 16 muestra el total de estudios relevantes por su tipo de publicación, teniendo como mayoría un total de 43 artículos científicos y con menor cantidad solamente 2 congresos.

Gráfica 16. Total de estudios por tipo de publicación

La Gráfica 17 muestra el total de estudios relevantes para herramientas por su tipo de publicación, teniendo como mayoría un total de 16 artículos científicos y con menor cantidad 1 capítulo de libro.

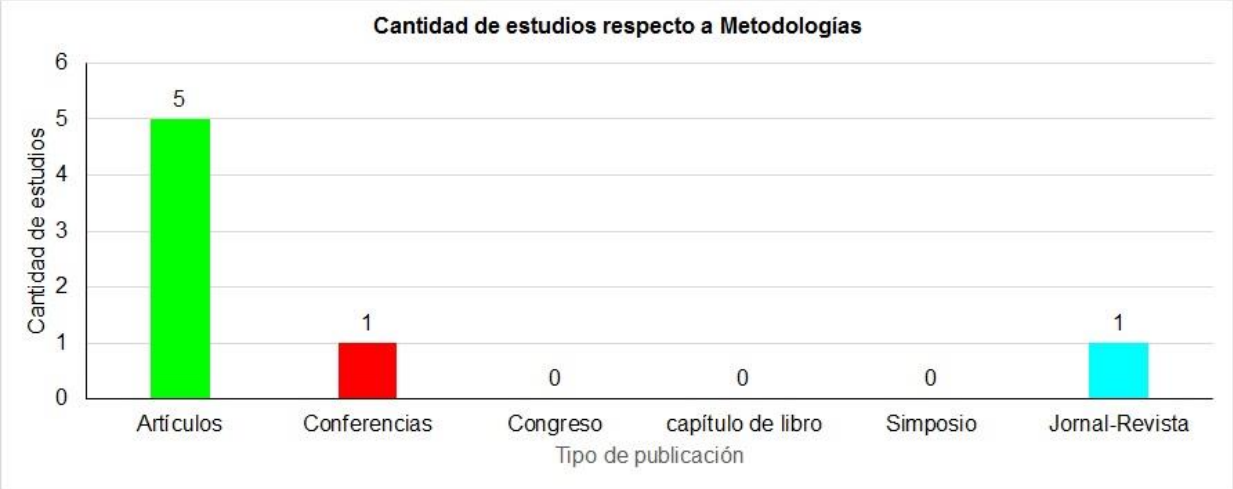
Gráfica 17. Total de herramientas por tipo de publicación

La Gráfica 18 muestra el total de estudios relevantes para frameworks por su tipo de publicación, teniendo como mayoría un total de 10 artículos científicos y con menor cantidad 1 Simposio.



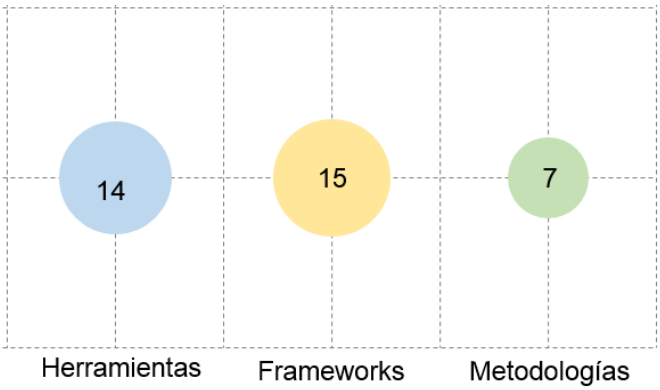
Gráfica 18. Total de frameworks por tipo de publicación

La Gráfica 19 muestra el total de estudios relevantes para metodologías por su tipo de publicación, teniendo como mayoría un total de 17 artículos científicos y con menor cantidad 2 capítulos de libros.



Gráfica 19. Total de metodologías por tipo de publicación

De acuerdo al análisis, clasificación y síntesis de la información, se llegó a un total de 36 estudios finales. La Gráfica 20 muestra el total de los estudios finales de los cuales 14 son herramientas, 15 frameworks y 7 Metodologías. En estos estudios se describen aportaciones en el procesamiento de la información para abordar el problema de la variedad en Big Data.



Gráfica 20. Total de Herramientas, Frameworks y Metodologías finales

5.2 Análisis de las preguntas de investigación

5.2.1 SQ1 = ¿Existe algún estudio de mapeo sistemático que aborde el problema de la variedad en Big Data?

Esta pregunta de investigación se centra en identificar y evidenciar los estudios que puedan evidenciar un estudio de mapeo sistemático para abordar el problema de la variedad por medio de alguna herramienta, framework o metodología. Dicho lo anterior, no se encontró algún estudio, mapa o mapeo sistemático que aborde el problema de la variedad por medio de Herramientas, IDE´s, frameworks o Metodologías.

La Tabla 6. Cadena y Búsqueda de estudios sin ajustes (ver Tabla 6), la cual muestra la cadena de búsqueda SQ1 "Búsqueda de los estudios de un mapa o mapeo sistemático en el problema de la variedad en sistemas Big Data".

SQ1= ((Title:(A systematic AND (mapping OR map) AND (in variety problem OR in variety) AND (Big Data OR Big Data systems))) OR (Abstract:(Systematic Mapping OR problem of variety OR variety in Big Data)) OR (Keywords:(Systematic Mapping OR Variety OR Variety problem OR Big Data systems OR Big Data)))

La cadena de búsqueda se ajustó a las diferentes fuentes, también la cadena de búsqueda fue diseñada para buscar estudios por "Titulo", "Resúmen" y "Palabras Clave". Como se mencionó anterior mente el resultado para un mapeo sistemático es de 0 estudios.

5.2.2 SQ2 = ¿Qué herramientas son empleadas para abordar el problema de la variedad en Big Data?

Esta pregunta de investigación tiene como objetivo demostrar la distribución de los artículos seleccionados y analizarlos para saber evidenciar las herramientas que ayuden a abordar el problema de la variedad en Big Data.

En un inicio se aplicó la cadena de búsqueda SQ2 "Búsqueda de estudios para herramientas, IDE´s que aborden el problema de la variedad en Big Data" sin ningún criterio de exclusión,

esto con el fin de tener un panorama general acerca de las herramientas utilizadas en Big Data dando como resultado un total de 79,599 estudios iniciales.
Como segunda fase, la cadena de búsqueda se ajustó de acuerdo a los criterios de inclusión por medio de las fuentes digitales, obteniendo un total de 5,175 estudios secundarios.

Por último, se ajustó los filtros manualmente a cada fuente digital con los criterios de inclusión exclusión, así como una discriminación manual de estudios que no cumplían con los criterios mencionados, obteniendo un total de 37 estudios Primarios.

SQ2 = ((Tools OR IDE) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))

La Tabla 22 muestra la lista completa de los estudios que abordan el problema de la variedad con diversas herramientas de análisis y procesamiento de datos.

Tabla 22. Lista del total de estudios mediante herramientas

Trabajos de Investigación	Artículos
Herramientas	[66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100], [101], [102].

Se llevó un análisis de la información de los 37 estudios primarios, donde se hace mención en la mayoría de los estudios que los datos no estructurados son uno de los mayores problemas en el análisis de Big Data y que también se identifican como "información humana" (Registros financieros, meteorológicos, imágenes, simulación, videos, etc.) y que el 80% de todos los datos generados se componen de datos no estructurados, que son aleatorios y no están modelados. (Sebastian, 2016)
Debido a que diariamente se genera una gran cantidad de datos no estructurados, es difícil y problemático simular y deducir los resultados de un análisis de información.

Algunas empresas han desarrollado herramientas que están enfocadas para el análisis y procesamiento de la información, como la Plataforma de análisis de Big Data de Amazon,

IBM InfoSphere BigInsights, Big Data Analytics de TERADATA, 1010data Big Data Platform, Cloudera Big Data Solution, entre otras. Estas compañías analizan gran cantidad de datos con la ayuda de diferentes tipos de herramientas y también proporcionan una interfaz de usuario fácil o simple para analizar datos.

Las herramientas que se encontraron para el análisis de los datos estructurados, semiestructurados y no estructurados se muestran en la Tabla 23.

Tabla 23. Lista de estudios Finales de Herramientas

#	Herramienta	Artículos
1	Hadoop	[66], [67], [70], [71], [73], [74], [75], [76], [77], [78], [79], [81], [83], [86], [87], [88], [89], [90], [91], [92], [94], [97].
2	Spark	[79], [74], [75], [77], [78], [79], [80], [84], [87], [88], [94], [97], [99].
3	Genus	[69].
4	IBM InfoSphere (Datastage)	[72], [74].
5	Storm	[66], [71], [73], [74], [75], [78], [81], [86], [87], [91], [92], [94], [98], [99].
6	IBM BigSheets	[72].
7	Project Voldemort	[68], [94], [99].
8	Cloudera (Hadoop)	[66], [74], [76], [81].
9	Flink	[73], [78], [81], [86], [87], [97], [99].
10	Samza	[73], [78], [87], [92], [99].
11	Mahout	[66], [74], [75], [78], [81], [86], [88], [91], [94], [99].
12	Hortonworks	[66], [72], [74], [83].
13	MapReduce	[66], [67], [68], [70], [71], [72], [73], [74], [76], [77], [78], [79], [80], [81], [86], [87], [89], [90], [92], [94], [97], [98], [99].
14	EpiC	[67].

La Tabla 24 muestra la clasificación de las herramientas por código abierto y de licencia.

Tabla 24. Clasificación de herramientas usadas en el análisis de la variedad en Big Data

Herramientas usadas en el análisis de la variedad de Big Data	
Herramientas de Código Abierto	**Herramientas de Licencia**
Apache Hadoop Apache MapReduce Apache Mahout Apache Storm Apache Spark Apache Flink Apache Samza Project Voldemort Hortonworks	EpiC Genus IBM InfoSphere IBM BigSheets Cloudera Big Data Solutions-(Hadoop)

La Tabla 25 muestra las principales características de las herramientas de licencia en el análisis de la información para tratar el problema de la variedad en Big Data.

Tabla 25. Herramientas de Licencia

Herramientas de licencia	
Nombre	**Características Principales**
EpiC	➢ Es un sistema extensible para abordar el desafío de la variedad de datos de Big Data. ➢ Manejar el desafío de la variedad de datos, por lo cual favorece una arquitectura híbrida. ➢ Desacopla el modelo de programación concurrente y el modelo de procesamiento de datos. ➢ Adopta un diseño extensible y soporta dos modelos de procesamiento de datos, MapReduce y el modelo de base de datos de relaciones. ➢ Utiliza MapReduce, administrando ciertos datos no estructurados de manera efectiva.

	➢ Divide el trabajo analítico de datos en sub tareas y elige los sistemas adecuados para realizar esas sub tareas en función de los tipos de datos. ➢ Resuelve el desafío del volumen de datos de Big Data mediante la paralelización.
Genus	➢ Es una nueva herramienta ETL (Extracción-Transformación-Carga) que trata el problema de la variedad. ➢ Con respecto a ETL que trata con la variedad de Big Data incluye varios formatos tales como datos de texto estructurados, numéricos, no estructurados. ➢ Extrae los datos de diferentes tipos de documentos: texto, imagen y video, los transforma y los carga en un almacén de datos de documentos. ➢ Implementa y valida en un estudio de caso comercial, con el propósito de ayudar a la inteligencia de Negocios (BI) en la toma de decisiones. ➢ El resultado de ésta herramienta propuesta es un almacén de datos mediante el lenguaje XML. ➢ Trabaja principalmente con texto, imágenes y el video.
IBM InfoSphere	➢ Es una plataforma de integración de datos permitiendo comprender, limpiar, supervisar y transformar datos. ➢ Se basa en Hadoop para mejorar sus capacidades y proporciona una interfaz de usuario interactiva para analizar gran cantidad de datos. ➢ Usa un Lenguaje de consulta declarativa para facilitar el análisis de información estructurada, no estructurada y semiestructurada. ➢ Ayuda a transformar información independientemente de su formato y entregarla a cualquier sistema, garantizando la agilidad en la generación de valor y la reducción del riesgo asociado a TI. ➢ Facilita la comprensión, de la información.
IBM BigSheets	➢ Es una herramienta de análisis y visualización. ➢ Utiliza una interfaz que permite analizar la cantidad de datos y los trabajos de recopilación de larga ejecución. ➢ Se utiliza para dividir grandes cantidades de datos no estructurados en contextos de negocios específicos para situaciones específicas.

	➢ Puede cargar datos de múltiples fuentes, como rastreadores web, bases de datos, archivos de texto, archivos Json, csv, etc. ➢ Puede almacenar los datos en el sistema de archivos de distribución para su procesamiento.
Cloudera Big Data Solutions-(Hadoop)	➢ Utiliza Apache hadoop para obtener una salida más valiosa de todos sus datos. ➢ Cloudera permite añadir las funciones de seguridad, control y gestión necesarias para establecer una base de nivel empresarial los datos. ➢ Funciones de almacenamiento de datos tradicionales para informes y modelos de Inteligencia de negocios. ➢ Análisis de datos para datos estructurados, registros de máquinas, texto y datos de IoT (Internet de las Cosas).

La Tabla 26 muestra las principales características de las herramientas de código abierto en el análisis de la información para tratar el problema de la variedad en Big Data.

Tabla 26. Herramientas de código abierto

Herramientas de Código Abierto	
Nombre	**Características Principales**
Apache Hadoop	➢ Proporciona seguridad y confiabilidad para el procesamiento de datos. ➢ Tiene su propio paradigma de cálculo de Hadoop llamado MapReduce, donde en el trabajo se divide en varias unidades y luego se procesa en un sistema agrupado o en una cuadrícula. ➢ Es un framework de procesamiento distribuido muy eficiente para procesar datos no estructurados, semiestructurados y estructurados.
Apache MapReduce	➢ Se relaciona con el Sistema de archivos distribuidos de Hadoop. ➢ MapReduce es que el sistema aborda el desafío del volumen ➢ Administra ciertos datos no estructurados de manera efectiva. ➢ MapReduce está reduciendo los datos redundantes y menos utilizados. ➢ Es buena para el procesamiento de texto
Apache Mahout	➢ Aprendizaje automático escalable y software de código abierto de minería de datos basado principalmente en Hadoop.

	➢ Extrae conocimiento útil a partir de fuentes de datos en bruto. ➢ Apache Mahout se apoya en la arquitectura Map-Reduce para realizar implementaciones escalables de diversos algoritmos clásicos de minería de datos y aprendizaje automático.
Apache Storm	➢ Es un sistema de código abierto distribuido en tiempo real. ➢ Storm hace en tiempo real el procesamiento de lo que Hadoop hizo para el procesamiento por lotes. ➢ Sirve para el procesamiento en tiempo real y análisis de flujo de datos. ➢ Procesa grandes conjuntos de datos estructurados y no estructurados en tiempo real.
Apache Spark	➢ Está integrado con Apache Hadoop. ➢ Trabaja en memoria, consiguiendo mayor velocidad de procesamiento. ➢ Permite trabajar en disco. ➢ Permite el procesamiento en tiempo real de los datos. ➢ Apache Spark permite a los programadores realizar operaciones sobre un gran volumen de datos en clústeres de forma rápida y con tolerancia a fallos.
Apache Flink	➢ Sirve para procesar datos tanto en modo de tiempo real como en modo de proceso por lotes. ➢ Cálculo tolerante a fallos y a gran escala. ➢ El modelo de programación de Flink es similar a MapReduce.
Apache Samza	➢ Realiza el procesamiento distribuido de stream utilizado en Kafka y YARN en el proceso. ➢ Cada tarea contiene un almacén clave-valor usado para almacenar el estado. ➢ No dispone de una librería con algoritmos de minería de datos. ➢ Alto rendimiento para analizar datos al instante. ➢ Es útil en casos en los que necesitamos procesar los datos en modo batch o modo streaming.

Project Voldemort	➢ Es un sistema de almacenamiento de valor-clave distribuido. ➢ Los datos se replican automáticamente en varios servidores. ➢ La replicación y colocación de datos se decide mediante una API simple. ➢ La capa de almacenamiento es completamente simulable.
Hortonworks	➢ Plataforma para el procesamiento de datos de múltiples cargas de trabajo a través de una variedad de métodos de procesamiento. ➢ Se utiliza para administración, integración, seguridad de operaciones.

La Tabla 27 muestra la comparación de las herramientas por formato, procesamiento y tipo de dato a tratar.

Tabla 27. Tabla comparativa de herramientas usadas en el procesamiento de Big Data

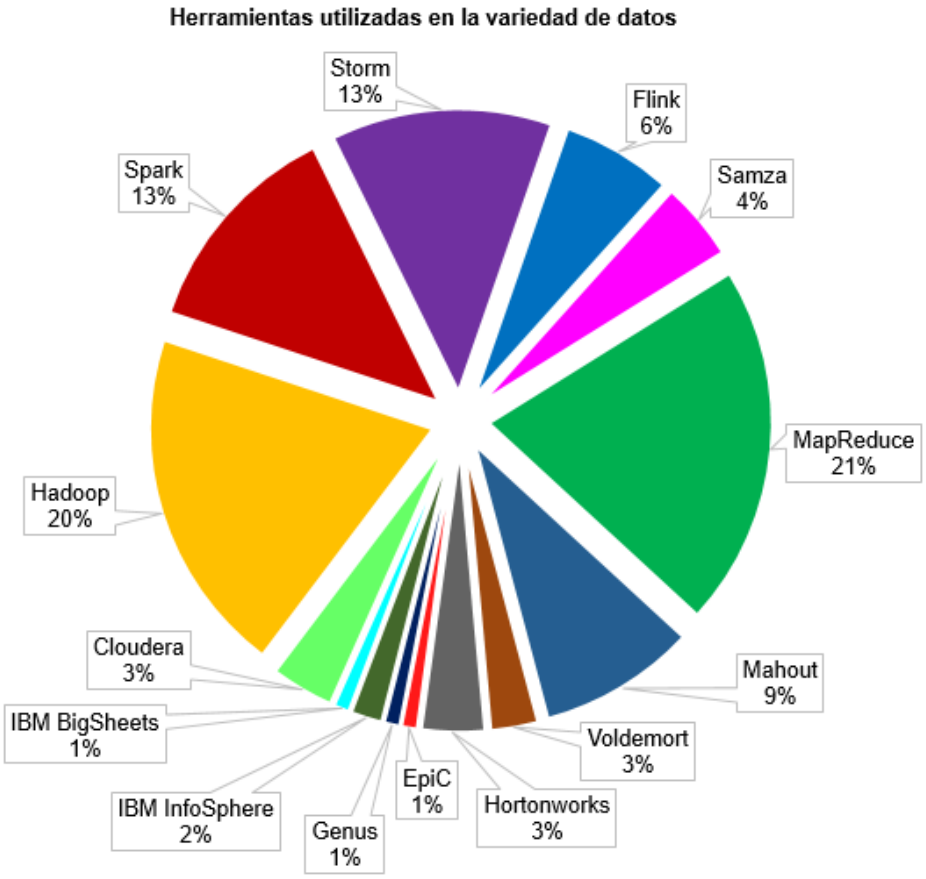
Acontinuación se muestran algunos acrónimos como referencia para la Tabla 27:

- RDD: Resilient Distributed Dataset (Conjunto de datos distribuidos resilientes).
- EST: Datos Estructurados.
- S-EST: Datos Semiestructurados.
- N-EST: Datos No estructurados.

Herramienta	**Formato de Datos**	**Procesamiento**	**Tolerancia a fallos**	**Tipo de Datos**
Hadoop	Valor - clave	Lotes	SI	EST, S-EST. y N-EST.
Spark	Valor – clave, RDD	Lotes y Stream	SI	EST, S-EST. y N-EST.
Storm	Valor - clave	Stream	SI	EST. y N-EST.
Flink	Valor - clave	Lotes y Stream	SI	Semiestructurados
Samza	Valor - clave	Stream	SI	Semiestructurados
MapReduce	Valor – clave, RDD	Stream	SI	No Estructurados
Mahout	Valor - clave	Lotes	No	Semiestructurados
Voldemort	Valor - clave	Lotes	No	S-EST. y N-EST.
Hortonworks	RDD	Lotes y Stream	SI	Semiestructurados
EpiC	Valor – clave, RDD	Lotes	SI	No Estructurados
Genus	Valor – clave, RDD	Lotes	SI	EST, S-EST. y N-EST.
IBM InfoSphere	Valor - clave	Lotes y Stream	SI	EST, S-EST. y N-EST.
IBM BigSheets	RDD	Lotes y Stream	SI	EST. y S-EST.
Cloudera	Valor - clave	Stream	SI	Estructurado

Como se puede observar, existen herramientas que nos ayudan a procesar los datos estructurados, semiestructurados y aquellos que no tienen estructura, por ejemplo: En la revisión de los estudios, la única herramienta que trata las imágenes y videos es Genus, proponiendo que el video se trabaje como una serie de imágenes y transformar los datos obtenidos en formatos XML para poder tener un formato semiestructurado. El texto, bases de datos, tablas y toda la información semiestructurada y estructurada son manejados en su mayoría por las herramientas de Apache (Hadoop, Mapreduce, Spark, Flink, etc.). Utilizando técnicas de optimización, modelos matemáticos, análisis semánticos, minería de datos, etc., y así poder aplicar técnicas de visualización para tener una mejor toma de decisiones.

En la Gráfica 21 se muestran las herramientas más usadas para el análisis de la información de datos estructurados, semiestructurados y no estructurados es Hadoop, MapReduce, Spark y Storm siendo estas herramientas de Apache.



Gráfica 21. Herramientas utilizadas en la variedad de datos

5.2.3 SQ3 = ¿Qué framework son utilizados para abordar el problema de la variedad en sistemas Big Data?

Esta pregunta de investigación tiene como objetivo demostrar la distribución de los artículos seleccionados y analizarlos para poder evidenciar los frameworks que ayudan a abordar el problema de la variedad en Big Data.

En un inicio se aplicó la cadena de búsqueda SQ3 "Búsqueda de estudios para frameworks" sin ningún criterio de exclusión, esto con el fin de tener un panorama más amplio acerca de los frameworks que se utilizan en el procesamiento de datos en Big Data, éste primer panorama arrojó un total de 2,278 estudios iniciales.
Como segunda fase, la cadena de búsqueda se ajustó de acuerdo a los criterios de inclusión por medio de las fuentes digitales, obteniendo un total de 540 estudios secundarios.

Por último, se ajustó los filtros manualmente a cada fuente digital con los criterios de inclusión exclusión, así como una discriminación manual de estudios que no cumplían con los criterios mencionados, obteniendo un total de 24 estudios Primarios.

SQ3 = ((Framework OR Frame) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))

La Tabla 28 muestra la lista completa de los estudios que abordan el problema de la variedad con diversos frameworks para el procesamiento de datos.

Tabla 28. Lista del total de estudios sobre frameworks

Trabajos de Investigación	Artículos
Frameworks	[103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126].

Se llevó un análisis de la información, donde se hace mención que la plataforma de código abierto Hadoop tiene el liderazgo en la actualidad como el mejor entorno para analizar grandes cantidades de datos. Ya que, está inspirado en el paradigma de programación

MapReduce, el cual consiste en dividir en dos tareas (mapa - reducir) logrando un alto paralelismo en el procesamiento.

Hadoop es el framework principal en la mayoría de los procesamientos de datos en Big Data por su facilidad de almacenar y analizar cantidades masivas de datos, tanto estructurados como sin estructurar.

En la actualidad existe un conjunto de frameworks para procesar los distintos tipos de datos con ayuda de Hadoop, ya que Hadoop se puede combinar perfectamente con distintas herramientas y plataformas. Por ejemplo, SAS incorpora Hadoop en sus aplicaciones. También SAS permite trabajar en memoria a través de Hadoop. IBM trabaja con Hadoop en su plataforma IBM InfoSphere BigInsights (BigInsights). Microsoft incluye Hadoop en SQL Server 2014, Windows Server 2012, HDInsight and Polybase. Oracle incluye Hadoop en Oracle Big Data Appliance, Oracle Big Data Connectors y Oracle Loader for Hadoop.
Los frameworks que se encontraron para el análisis de los diferentes tipos de datos se muestran en la Tabla 29.

Tabla 29. Lista de estudios Finales de frameworks

#	Framework	Artículos
1	Marimba	[103]
2	Framework for Unstructured Data Analysis	[104]
3	Analytics-as-a-Service (AaaS)	[105]
4	RUBA	[106]
5	Framework for Concept-Based Exploration of Semi-Structured Software Engineering Data	[107]
6	Framework to Handle Data Heterogeneity Contextual	[108]
7	Framework ETL	[109]
8	Framework of Integrated Big Data	[110]
9	jMetalSP	[111]
10	Project assessment Framework (BigDAF)	[112]

11	Twister.Net	[113], [115]
12	Dryad y DryadLINQ	[113]
13	Nephele	[113]
14	Piccolo	[113]
15	Framework for Extracting Reliable Information from Unstructured Uncertain Big Data	[114]

La Tabla 30 muestra un resumen de las características principales de los principales frameworks para el análisis y procesamiento de datos.

Tabla 30. Características principales de los frameworks

Nombre	Características Principales
[103] Marimba	➢ Basado en Hadoop y se puede usar para implementar trabajos MapReduce. ➢ Se puede ejecutar sobre cualquier versión de Hadoop. ➢ Trabajos de MapReduce de manera incremental mediante la recalculación de datos. ➢ Definición una vista materializada mediante una consulta declarativa. ➢ Su funcionamiento principal es por medio de la detección de Deltas "Δ" y por recuento incremental de palabras. ➢ Utiliza algoritmos de cálculo llamados: TextLongMapAbelianis y Friends of Friends.
[108] A Framework to Handle Data Heterogeneity	➢ Es un framework prototipo. ➢ Mejora el método de transformación de datos en una base de datos No-SQL o H-Base. ➢ Procesa datos estructurados centralizados a datos estructurados distribuidos ➢ Procesa datos no estructurados a un formato estructurado ➢ Procesa datos semiestructurados a un formato estructurado ➢ Utiliza el entorno basado en Hadoop y H-Base.

	➢ Almacena y recupera conjuntos de datos heterogéneos. ➢ Desarrollaron un algoritmo para almacenar el formato de datos no estructurados. ➢ A los datos no estructurados se le da una forma estructurada utilizando MapReduce. ➢ Conversión de los datos no estructurados a un formato semiestructurado XML para almacenarlos en H-Base.
[110] Framework of Integrated Big Data	➢ Framework integrado mediante: el almacenamiento, la gestión de datos, el análisis y la visualización de datos. ➢ Unifica los datos estructurados, semi y no estructurados en un espacio idéntico. ➢ Realiza operaciones de inserción, eliminación, actualización y consulta de datos. ➢ La computación integrada se ocupa de los tres tipos de datos al mismo tiempo. ➢ Se tiene una visualización unificada de los diferentes tipos de datos
[112] Project assessment Framework (BigDAF)	➢ Framework llamado BigDAF ➢ Evalua la complejidad de los proyectos de Big Data. ➢ Ayuda a encontrar un clúster que defina mejor el "problema" de acuerdo con las dimensiones de las 3v. ➢ BigDAF proporciona cuatro resultados de evaluación distintos: • [100-200] Problema tradicional de BI • 200-300] Problema de BI cerca del desafío Big Data • [300-400] Problema de Big Data • [400-500] Problema complejo de Big Data. ➢ Menciona que se debe hacer dependiendo de la evaluación de los problemas, desde el uso de herramientas hasta el análisis de datos.

[113], [114], [115] Frameworks in Big Data Era	➢ Procesamiento eficiente de datos paralelos y distribuidos. ➢ Mencionan varios lenguajes de programación para el análisis de datos ➢ Muestran un análisis de los frameworks más utilizados para el análisis de Big Data: • MapReduce (MR). • Hadoop. • Twister. • Dryad. • DryadLINQ. • Nephele. • Piccolo.
[105] Analytics-as-a-Service (AaaS)	➢ Framework AaaS que extrae información de datos no estructurados. ➢ Análisis de datos mediante bases de datos NoSQL y contenido textual. ➢ Uso de algoritmos para llevar un filtrado y etiquetado de datos. ➢ Abarca tres actores principales: los usuarios, el frontend y las fuentes de datos. ➢ El usuario especifica las fuentes de datos para extraer los datos. ➢ Uso de un motor semántico para una serie de revisiones para mejorar la calidad de los datos obtenidos. ➢ Visualización por medio del navegador en forma de tablas y como panel de control.
[106] RUBA	➢ Framework propuesto para el análisis de datos no estructurados. ➢ Utiliza un motor de procesamiento de eventos complejos para analizar datos en tiempo real. ➢ Conviene los datos no estructurados en datos estructurados. ➢ Tiene cuatro módulos de procesamiento:

	• Módulo de Recepción de datos • Módulo de envío de datos • Módulo de procesamiento de datos no estructurados • Módulo de análisis en tiempo real. ➢ Tiene un flujo de datos desde la recopilación de datos hasta el envío de los resultados al usuario.
[107] A Generic Framework	➢ Framework propuesto (no desarrollado) para la exploración y consulta de datos semiestructurados. ➢ Uso de una red conceptual a partir de un contexto formal y un conjunto de datos. ➢ La construcción de la red conceptual estará formada por objetos y atributos. ➢ Los objetos son un conjunto de documentos. ➢ Los atributos son palabras clave extraídas de los documentos. ➢ Propone una interfaz de nube de etiquetas. ➢ Las nubes de etiquetas son un método de visualización para datos textuales. ➢ La navegación de la red es mediante un conjunto de palabras clave contenidas en los documentos de un conjunto de datos.
[124] A Framework for Unstructured Data Analysis	➢ Framework en desarrollo ➢ La idea inicial es un almacén de datos no estructurados ➢ La obtención de datos será por medio de tweets públicos ➢ Usará Hbase para almacenar los datos después de un análisis. ➢ Utilizará HBase mediante llamadas REST ➢ Procesamiento de datos mediante el clúster Hadoop. ➢ Utilizará Java para interactuar con el servidor y posteriormente se obtendrá un XML.
[109] Semantic Framework ETL	➢ Framework semántico ETL para la integración de datos. ➢ Utilizan tecnologías semánticas para conectar, vincular y cargar datos en un almacén de datos.

	➢ Modelo de datos semántico a través de ontologías. ➢ Creación de datos semánticos integrados utilizando RDF como modelo de datos gráficos. ➢ Utilizan SPARQL como lenguaje de consulta semántica. ➢ Extracción de datos en formatos de archivo plano. ➢ Transformación: Implica la limpieza de datos. ➢ Carga: Implica la propagación de los datos en un almacén de datos.
[111] Framework Algorítmico guiado por semántica	➢ Framework llamado jMetalSP ➢ Es una librería de algoritmos de optimización adaptados para Big Data ➢ Optimiza los problemas en de Big Data en tiempo real. ➢ Desarrollo de una ontología llamada BIGOWL ➢ Optimiza el procesamiento de Apache Spark ➢ Aborda los problemas de procesamiento en streaming ➢ Actúa ante cambios en el problema, datos o entorno (Variabilidad) ➢ Acerca el proceso de optimización (Valor, Veracidad).
[114] A Framework for Extracting Reliable Information	➢ Prototipo de la creación de un framework ➢ Su objetivo es extraer información confiable de datos inciertos no estructurados. ➢ Pretende desarrollar nuevos modelos para apoyar la toma de decisiones. ➢ Se proponen dos enfoques para manejar los tipos de datos no estructurados e inciertos. ➢ La forma principal es utilizar un enfoque de combinación de información ➢ La segunda forma es a través de la lógica difusa, enfoques de conjuntos suaves y técnicas de optimización robustas. ➢ Proponen los siguientes pasos de análisis: 1. Selección de fuente de Big Data no estructurada

	2. Filtrado de datos innecesarios 3. Modelo para extraer información confiable y no confiable 4. Tomar una decisión basada en la información extraída. ➢ Para lograr la extracción y procesamiento de datos es mediante Hadoop.
[Extra] Hadoop, MapReduce	➢ Utilizados para el análisis de datos mediante la extracción del conocimiento. ➢ Permiten la división de tareas muy intensas de cómputo masivo. ➢ Hadoop está compuesto de tres piezas: • Hadoop Distributed File System (HDFS) • Hadoop MapReduce • Hadoop Common. ➢ Los datos en el cluster de Hadoop son divididos. ➢ Las funciones map y reduce pueden ser ejecutadas en pequeños subconjuntos. ➢ MapReduce es el núcleo de Hadoop. ➢ El proceso es map separados los elementos en (clave/valor). ➢ El proceso reduce obtiene la salida de map en un conjunto más pequeño. ➢ Hadoop Common Components son un conjunto de librerías de varios sub proyectos de Hadoop ➢ Hadoop incorpora las herrmientas: • HBase, Jaql, Lucene, Pig • Mahout, Hive. ➢ Otras herramientas que implementan hadoop son: • IBM Power Systems • Windows Server y Windows Azure basadas en: Apache Fladoop ➢ SAS/ACCESS Interface to Hadoop.

La Tabla 31 muestra la comparación de las herramientas por formato, procesamiento y tipo de dato a tratar.

Tabla 31. Tabla comparativa de frameworks usados en el procesamiento de datos en Big Data

Acontinuación se muestran algunos acrónimos como referencia para la Tabla 31:

- EST: Datos Estructurados.
- S-EST: Datos Semiestructurados.
- N-EST: Datos No estructurados.
- HDFS: Hadoop Distributed File System

Framework	**Procesamiento**	**Formato de datos**	**Herramienta - IDE**	**Tipo de datos**
Marimba	Streaming,	Valor-Clave	MapReduce	EST, S-EST, y N-EST.
Framework for Unstructured Data Analysis	Lotes y Streaming	Valor-Clave	HBase, Hadoop	No estructurados
Framework (AaaS)	Lotes - Algoritmos	Valor-Clave	MapReduce, Hadoop, Mahout y CouchDB.	No estructurados
RUBA	Streaming	.	CQL, RUBA processor y Motor CEP (Complex Event processing)	N-EST y EST
Framework for Concept-Based Exploration of Semi-Structured Software Engineering Data	Red conceptual – nube de etiquetas	Valor - Clave	Prototipo (MapReduce)	Semiestructurados
Framework to Handle Data Heterogeneity Contextual	Lotes-Algoritmos	Valor-Clave	No-SQL, HBase, Hadoop, MapReduce	EST, S-EST, y N-EST.

Framework ETL	Streaming	Semántico	Framework de Descripción de Recursos (RDF), SPARQL	No estructurados
Framework of Integrated Big Data	Lotes	Valor- Clave	Voldemort, Redis, Hypertable, HBase, MongoBD, CouchBD, Neo4j, Graph y Tensor	EST, S-EST, y N-EST.
jMetalSP	Streaming-Algoritmos	Semántico	jMetal, Apache Spark , BIGOWL	No estructurados
BigDAF	Streaming	Semántico	Hadoop, BigDAF	EST, S-EST, y N-EST.
Twister.Net	Streaming	Valor - Clave	MapReduce	Estructurados
Dryad y DryadLINQ	Streaming	Valor-Clave	Hadoop	EST, S-EST, y N-EST.
Nephele	Streaming	Valor-Clave	Hadoop, MapReduce, Dyrad, HDFS	EST, S-EST, y N-EST.
Piccolo	Streaming	Valor-Clave	------	No estructurados
Framework for Extracting Reliable Information from Unstructured Uncertain Big Data	Streaming	Valor - Clave	Prototipo (Hadoop)	No estructurados

Como se muestra en la Figura 5, los elementos principales son los datos obtenidos a partir de cualquier fuente de información, a menudo en múltiples formatos. Todos estos datos deben agruparse para fines de procesamiento y análisis.

Los datos inicialmente están en estado sin procesar y necesitan ser transformados, para ello se necesitan procesar de distintas maneras dependiendo del tipo de dato.

Cuando se obtiene un tipo de dato limpio y agrupado es necesario darle un formato, para ello los datos llevan un proceso de transformación por medio de distintos tipos de frameworks para después poder extraer la información más relevante para su posterior uso.

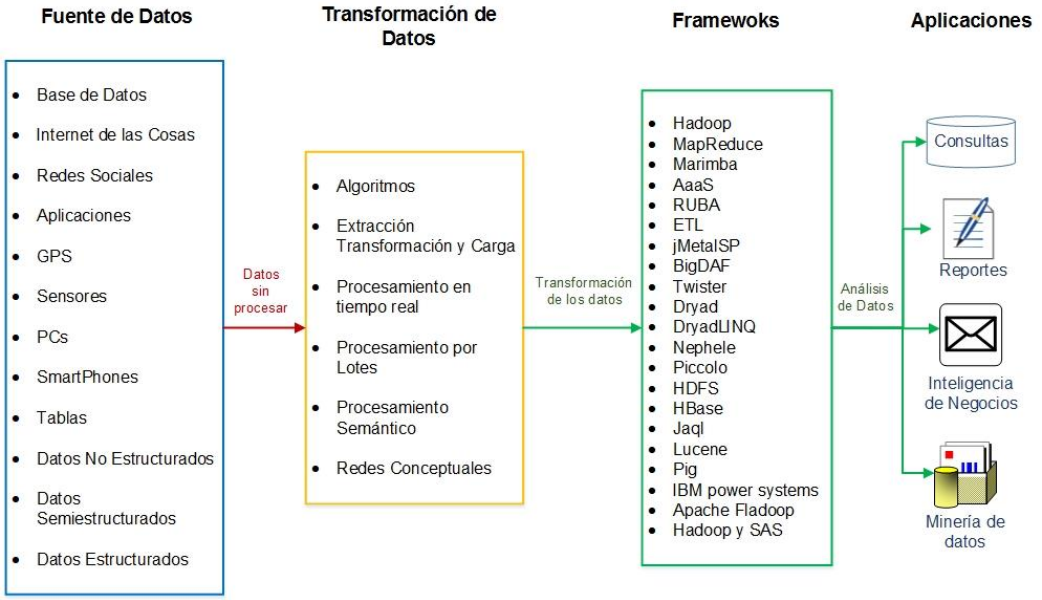


Figura 5. Extracción, Transformación y Análisis de los datos de Big Data

A pesar de que los frameworks utilizan herramientas, algoritmos y modelos similares, el procesamiento de los datos es completamente diferente.

5.2.4 SQ4 = ¿Qué Metodologías son empleadas para abordar el problema de la variedad en sistemas Big Data?

Esta pregunta de investigación tiene como objetivo demostrar la distribución de los artículos seleccionados y analizarlos para saber si existe alguna metodología empleada o desarrollada que ayude a resolver el problema de la variedad en Big Data.

En un inicio se aplicó la cadena de búsqueda SQ4 "Búsqueda de estudios para Metodologías" sin ningún criterio de exclusión, esto con el fin de tener un panorama más amplio acerca de los métodos o metodologías que se utilizan en el procesamiento de datos en Big Data, éste primer panorama arrojó un total de 4,086 estudios iniciales.
Como segunda fase, la cadena de búsqueda se ajustó de acuerdo a los criterios de inclusión por medio de las fuentes digitales, obteniendo un total de 415 estudios secundarios.

Por último, se ajustó los filtros manualmente a cada fuente digital con los criterios de inclusión exclusión, así como una discriminación manual de estudios que no cumplían con los criterios mencionados, obteniendo un total de 27 estudios Primarios.

SQ4 = ((Methodology OR Method) AND (variety OR variety problem) AND (Big Data OR Big Data Systems))

La Tabla 32 muestra la lista completa de los estudios que abordan el problema de la variedad con diversas metodologías, métodos y algoritmos para el procesamiento de datos.

Tabla 32. Lista del total de estudios sobre Metodologías

Trabajos de Investigación	Artículos
Metodologías	[127], [128], [129], [130], [131], [132], [133], [134], [135], [136], [137], [138], [139], [140], [141], [142], [143], [144], [145], [146], [147], [148], [149], [150], [151], [152], [153].

Se llevó un análisis de la información, donde se hace mención que no hay una Metodología definida para llevar a cabo el problema de la variedad de los datos, pero se pueden utilizar varios modelos y métodos para tratar de abordar este problema.

Las metodologías que se encontraron para el análisis de datos se muestran en la Tabla 33.

Tabla 33. Lista de estudios Finales de Metodologías

#	Metodología	Artículos
1	An Architecture and Methods for Big Data Analysis	[127]
2	Big Data: Methods, Prospects, Techniques.	[128]
3	A Storage Model for Handling Big Data Variety	[129]
4	Multi-model Databases: A New Journey to Handle the Variety of Data	[130]
5	Beyond the hype: Big Data concepts, methods and analytics	[131]
6	Big Data preprocessing: methods and prospects (IMAGS)	[132]
7	An Iterative Methodology for Big Data Management, Analysis and Visualization	[133]

La Tabla 34 muestra un resumen de las características principales de las principales Metodologías para el análisis de datos.

Tabla 34. Resumen de las características principales de las Metodologías

Acontinuación se muestran algunos acrónimos como referencia para la Tabla 34:

- EST: Datos Estructurados.
- N-EST: Datos No estructurados.
- S-EST: Datos Semiestructurados.

Artículo	Aportación	Métodos/Modelos	Objetivos	Procesamiento	Tipos de datos
An Architecture and Methods for Big Data Analysis	Método	•Arquitectura basada en la nube •Hadoop	Procesar e interpretar datos estructurados, no estructurados y semiestructurados, con el fin de proporcionar una plataforma que pueda extraer información de grandes fuentes de datos.	•Indexación y clasificación de datos •Minería de datos •Integración y unificación de datos •Análisis y limpieza de datos •Escalabilidad y elasticidad; •Control de la calidad de los resultados.	EST, S-EST y N-EST.
Big Data: Methods, Prospects, Techniques.	Métodos	•Métodos de minería de Big Data •Computación en la nube. •MapReduce.	Presentar métodos y técnicas de Big Data para analizar los distintos tipos de datos	•Procesamiento por lotes. •Paralelismo basado en datos distribuidos •Hashing •Indexación •Filtro de floración	EST, S-EST y N-EST.

		•Métodos de procesamiento basado en Hadoop		•Computación paralela.	
Multi-model Databases: A New Journey to Handle the Variety of Data	Modelos	•Modelo relacional •Modelo semiestructurado para documentos XML y JSON •Modelo de clave / valor •Modelo de grafos	Presentar multimodelos para crear una plataforma de base de datos para administrar la diversidad de los tipos de datos.	N/A	EST, S-EST y N-EST.
An Iterative Methodology for Big Data Management, Analysis and Visualization	Metodología	Bases de datos NoSQL RDBMS extendido Hadoop / MapReduce.	Metodología para abordar proyectos de Big Data de manera sistemática	•Etapas de datos Adquisición y gestión de fuentes de datos • Agregar valor a los datos • Selección e implementación de un Big Data Warehouse •Desarrollo de visualizaciones para Big Data.	S-EST y N-EST.

A Storage Model for Handling Big Data Variety	Modelo	Algoritmos de organización y procesamiento de la información en formato XML	Propuesta de modelo de almacenamiento basado en XML para resolver el problema de almacenar cualquier tipo de datos.	Uso de bases de datos no relacionales: Cassandra, SQL, MongoDB, Neo4j, HDFS.	No Estructurados.
Beyond the hype: Big Data concepts, methods, and analytics	Métodos	•Análisis de texto •Análisis de audio •Analítica de video •Analítica predictiva	Describir los métodos analíticos más utilizados para el análisis de Big Data	• Técnicas y algoritmos de extracción de información. • Enfoque basado en la transcripción •Enfoque basado en la fonética	No Estructurados
Big Data preprocessing: methods and prospects (IMAGS)	Métodos	• TF-IDF • Word2Vec • CountVectorizer • Tokenizer •StopWordsRemover • n-gram	Mostrar los métodos y técnicas de pre procesamiento de datos para la minería de datos en Big Data.	Pre procesamiento de datos. Transformación de datos, integración, limpieza y normalización.	No Estructurados

Durante el análisis de la información no se encontró alguna metodología que se enfoque únicamente al tratamiento de los diferentes tipos de datos en Big Data, por lo cual una alternativa para llevar a cabo una metodología para abordar el problema de la variedad de datos en Big Data sería implementar un grupo de modelos de acuerdo a las características presentadas a continuación:

- Visualización de los datos

  La visualización de los datos se ha convertido en uno de los métodos más aprovechados debido a su facilidad de interpretar y detectar patrones en los datos. (T. Roberto, 2015)

- Análisis de series temporales

  Consiste de un conjunto de técnicas estadísticas que analizan secuencias de datos para predecir la probabilidad de un resultado. La analítica predictiva se establece y aprende del pasado para proyectar determinados escenarios en un futuro, esto para la ayuda en la toma de decisiones. (H. Muhammad, 2016)

- Análisis de textos

  La minería de textos consta en técnicas de análisis de la información mediante procesos semánticos de grandes volúmenes de textos. Esto con el fin de extraer valor de la información procesada para un determinado fin. (G. Amir Gandomi, 2015)

- Algoritmos de procesamiento

  Se trata de algoritmos capaces de detectar una relación entre distintas variables de procesamiento de datos en una variedad de distintas bases de datos. Esto con el fin de estructurar diversos tipos de datos y agruparlos en bases de datos unificados. (I. Bogdan, 2018)

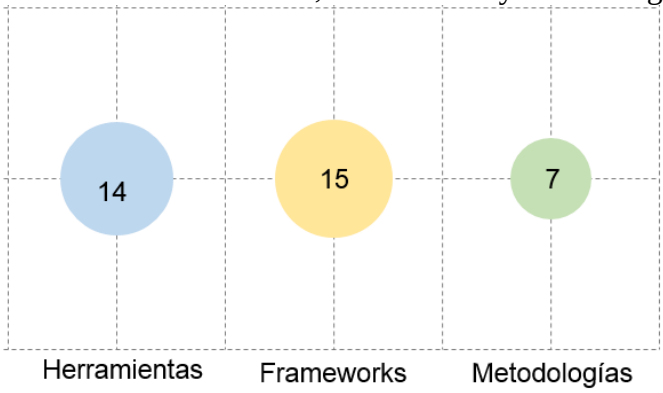
Capítulo 6

Principales hallazgos

6.1 Aspectos Relevantes

Retomando la Gráfica 19 se obtuvieron un total de 36 estudios relevantes de los cuales 14 son herramientas, 15 Frameworks y 7 Metodologías. En estos estudios se describen aportaciones en el procesamiento de la información para abordar el problema de la variedad en Big Data.

Gráfica 20. Total de Herramientas, Frameworks y Metodologías finales



La mayoría de las herramientas y frameworks tienen como principio fundamental el procesamiento mediante Hadoop y MapReduce.

La importancia de Hadoop radica básicamente en que permite desarrollar tareas muy intensivas de computación masiva o mejor llamado procesamiento por lotes, dividiéndolas en pequeñas tareas y distribuyéndolas en un conjunto de máquinas.

Hadoop está basado en el paradigma de programación MapReduce, el cual consiste en dividir en dos tareas (map - reduce) la manipulación de los datos distribuidos a nodos de un clúster (conjunto de máquinas) logrando un alto paralelismo en el procesamiento.

De acuerdo a lo anterior se puede entender que MapReduce es un distribuidor de tareas que permite de una forma bastante sencilla el repartir un conjunto pequeño de tareas entre un grupo de clúster, por lo tanto, se puede decir que MapReduce es el núcleo de Hadoop. (D. Jeff, 2004)

El procesamiento que realiza MapReduce en realidad se refiere a dos procesos separados que Hadoop ejecuta:

El proceso Map: este proceso toma un conjunto de datos y lo convierte en otro conjunto, donde los elementos individuales son separados en tuplas (pares de clave/valor).

El proceso Reduce: este proceso obtiene la salida del map como datos de entrada y combina las tuplas en un conjunto más pequeño de las mismas. La Figura 6 muestra el proceso típico de MapReduce.

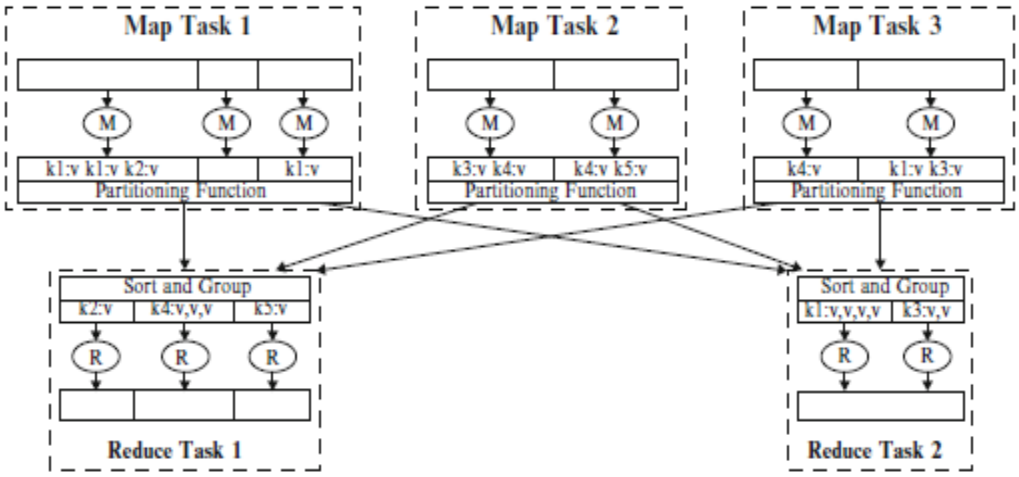


Figura 6. Procesamiento distribuido mediante MapReduce tomado de (D. Jeff, 2004)

Durante el análisis de las metodologías encontradas en el proceso del mapeo sistemático, no se encontró una implementación de alguna metodología que pueda abordar el problema de la variedad como tal, pero si se obtuvieron estudios que proponen modelos para abordae el problema mediante el procesamiento de la información con diversos métodos de almacenamiento y el análisis de la información por medio de algoritmos matemáticos.

De acuerdo al mapeo sistemático desarrollado se obtuvieron una serie de herramientas, frameworks y metodologías que abordan el problema de la variedad en Big Data. La Tabla 35 muestra las características principales de las herramientas, frameworks y metodologías.

A continuación, se muestra la Tabla 35 donde se describen las principales características de las Herramientas, Frameworks y Metodologías para procesar y analizar los diferentes tipos de datos.

Tabla 35. Resumen de las características principales de las Herramientas, Frameworks y Metodologías

Acontinuación se muestran algunos acrónimos como referencia para la Tabla 34:

- EST: Datos Estructurados.
- S-EST: Datos Semiestructurados.
- N-EST: Datos No estructurados.
- N/A: No Aplica

Herramienta		**Tipo de Datos**			**Procesamiento**			**Entrada**	**Salida**	**Tecnologías que utiliza**
		EST	**S-EST**	**N-EST**	**Lotes**	**Stream**	**Otro**			
1	Hadoop	**X**	**X**	**X**	**X**			Texto CSV JSON TSV Snappy Gzip Deflate Bzip2 Datos en formato binario, formato basado en filas, formato basado en columnas, etc.	Archivo binario Almacén de datos Clave - Valor	HDFS MapReduce

2	Spark	**X**	**X**	**X**	**X**	**X**		Texto CSV JSON Parquet Kafka Archivos de secuencia HDFS HBase Cassandra	Csv RDD	Hadoop MapReduce
3	Genus	**X**	**X**	**X**	**X**			Texto Imagen Video CSV JSON XML HBase	XML en un almacén de datos NoSql.	Algoritmos de ETL
4	IBM InfoSphere (Datastage)	**X**	**X**	**X**	**X**	**X**		Texto CSV XML ZIP Datos en formato binario	CSV y Formato tipo binario	Basado en Hadoop Algoritmos ETL

								Tipos de datos numéricos		
5	Storm	**X**		**X**		**X**		Texto Tuplas de datos	Texto, XML	Apache ZooKeeper
6	IBM BigSheets	**X**	**X**		**X**	**X**		CSV JSON TSV Hojas de datos.	Interfaz gráfica similar a la hoja de excel	
7	Project Voldemort		**X**	**X**	**X**			Texto JSON XML	sistema de almacenamiento de clave - valor	Hadoop
8	Cloudera		**X**			**X**		CSV JSON Datos en formato binario	Archivo en formato binario y Almacen de datos por clave y valor	Hadoop Apache Sqoop Apache Flume Apache Kafka Apache Hive Apache Pig MR2 Apache Spark
9	Flink			**X**	**X**	**X**		Texto Aplicaciones, Sensores, Dispositivos, Archivos de sistema, Registros de mensajes, Redes sociales.	Almacenamiento Clave – valor en HDFS.	YARN, Mesos, Kubernetes, Docker o standalone, Hadoop, HDFS, YARN, HBase
10	Samza		**X**			**X**		Procesamiento y transformación de datos desde cualquier fuente	Almacenamiento Clave-valor	Apache Kafka, AWS Kinesis , Azure EventHubs ,

								semiestructurada, incluido Apache Kafka.		ElasticSearch y Apache Hadoop.
11	Mahout			**X**	**X**			Foursquare como Motor de recomendaciones. Redes sociales para Modelado de intereses de usuario. Yahoo! para Minería de patrones.	Archivos de vectores Mahout y un diccionario de término que luego se podrán utilizar para un agrupamiento en forma clave-valor	MapReduce, kmeans, fuzzy k-means, Canopy, Dirichlet, Mean-Shift.
12	Hortonwork			**X**	**X**	**X**		Texto procesamiento y transformación de datos desde cualquier fuente	Almacenamiento en HDFS y HBase	Apache Hadoop, HDFS, MapReduce, Pig, Hive, HBase, NiFi y Spark.
13	MapReduce			**X**		**X**		Texto Datos en forma de archivos o directorios	tuplas (pares clave/valor) almacenados en HDFS	Hadoop, MapReduce
14	EpiC			**X**	**X**			Texto Correo electrónico,	Grafos por Búsqueda en profundidad,	Hadoop, MapReduce, Pregel, Dryad, Graph

								Registro telefónico.	Almacén Clave valor mediante base de datos distribuida	
Frameworks		**EST**	**S-EST**	**N-EST**	**Lotes**	**Stream**	**Otro**	**Entrada**	**Salida**	**Tecnologías**
15	Marimba	**X**	**X**	**X**		**X**		Texto Datos gráficos (PowerPoint) Sitios web, Publicaciones de blog CSV JSON XML HBase	Formato clave-valor (deltas), HBase	MapReduce, Hadoop, HBase
16	Framework to Handle Data Heterogeneity Contextual	**X**	**X**	**X**	**X**		Algoritmos	Texto tweets públicos de Twitter XML HBase	Almacenamiento en HBase usando el cluster de Hadoop	Hadoop, HBase
17	Framework of Integrated Big Data	**X**	**X**	**X**	**X**			Texto Audio Imágenes Datos gráficos XML	Modelo de almacenamieto clave-valor en una BD NoSQL con formatos	Graph, tensor, HBase, Hadoop, MapReduce

								HTML Tablas	XML, YAML, y JSON.	
18	BigDAF	**X**	**X**	**X**		**X**		Texto Web-log (blogs) Sensores Archivos de dispositivos JSON XML HBase	Almacén de datos en HDFS	Algoritmos, Hadoop, HDFS
19	Dryad y DryadLINQ	**X**	**X**	**X**		**X**		Texto XML HTML Cassandra HBase	Almacén de datos clave-valor	Hadoop
20	Nephele	**X**	**X**	**X**		**X**		Procesamiento en la nube CSV, XML, JSON, Tupas de bases de datos	Almacén de datos clave-valor	Hadoop, MapReduce, Dyrad, HDFS
21	Framework (AaaS)	**X**	**X**	**X**	**X**		Algoritmos	Contenido Textual	salida de datos en interfaz,	MapReduce, Hadoop, Mahout, CouchDB,

								Formato Clave – valor XML HTML RTF Bases de datos: Oracle DB2	Almacenamiento en formato clave-valor.	Oracle, DB2, R, Cluto, Weka.
22	RUBA	**X**		**X**		**X**		CCTV, Sistemas de monitorización, Imágenes, video. Bases de datos CEP	datos almacenados en una base de datos CEP(complex event processing) utilizando CQL(continuous query language)	Algoritmos recording, rebuilding and re-execution.
23	Twister.Net	**X**				**X**		Procesamiento en la nube: Bases de datos, Tuplas, Hojas de cálculo	Almacén de datos clave-valor	Hadoop, MapReduce,
24	Framework for Concept-Based		**X**				Red Conceptual / Nube de etiquetas	JSON XML HTML, .Java	archivos clave-valor con información	algoritmos por medio de nube de etiquetas, Mapreduce

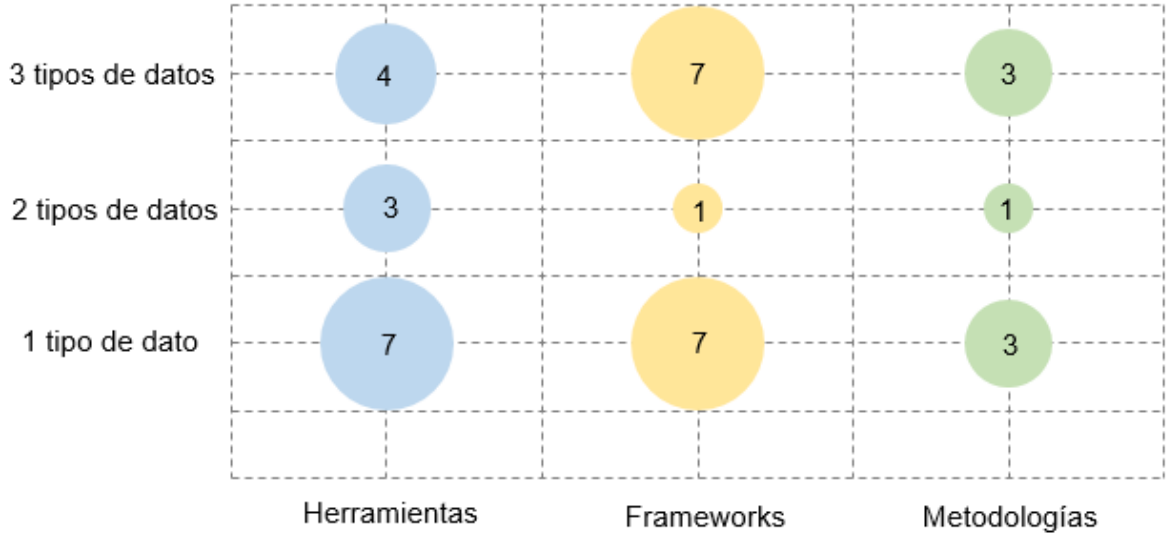
	Exploration of Semi-Structured Software Engineering Data								basada en etiquetas	
25	Framework for Unstructured Data Analysis			**X**	**X**	**X**		Texto y tweets de Twitter	almacenamiento de datos en formato clave-valor en Hbase	HBase, Hadoop, MapReduce, Apache Hive
26	Framework ETL			**X**		**X**		Semántico, web	Framework de descripción de recursos (RDF) como modelo de datos gráficos	Framework de Descripción de Recursos (RDF), SPARQL
27	jMetalSP			**X**		**X**	Algoritmos	Web Semántica	librería de algoritmos de optimización	jMetal, Apache Spark, BIGOWL
28	Piccolo			**X**		**X**		Texto sitios web, datos gráficos	Almacén de datos clave-valor	-----
29	Framework for Extracting Reliable Information from Unstructured Uncertain Big Data			**X**		**X**		Texto Web, Audio	Almacén de datos clave-valor	Hadoop

Metodologías		EST	S-EST	N-EST	Método	Modelo	Metodología	Objetivos	Procesamiento de datos	Tecnologías que utiliza
30	An Architecture and Methods for Big Data Analysis	**X**	**X**	**X**	**X**			Procesar e interpretar datos estructurados, no estructurados y semiestructurados, con el fin de proporcionar una plataforma que pueda extraer información de grandes fuentes de datos.	•Indexación y clasificación de datos •Minería de datos •Integración y unificación de datos •Análisis y limpieza de datos •Escalabilidad y elasticidad; •Control de la calidad de los resultados.	•Arquitectura basada en la nube •Hadoop
31	Big Data: Methods, Prospects, Techniques	**X**	**X**	**X**	**X**			Presentar métodos y técnicas de Big Data para analizar los distintos tipos de datos	•Procesamiento por lotes. •Paralelismo basado en datos distribuidos •Hashing •Indexación •Filtro de floración •Computación paralela.	•Métodos de minería de Big Data •Computación en la nube. •MapReduce. •Métodos de procesamiento basado en Hadoop
32	Multi-model	**X**	**X**	**X**		**X**		Presentar multimodelos	N/A	•Modelo relacional

	Databases: A New Journey to Handle the Variety of Data							para crear una plataforma de base de datos para administrar la diversidad de los tipos de datos.		•Modelo semiestructurado para documentos XML y JSON •Modelo de clave / valor •Modelo de grafos
33	An Iterative Methodology for Big Data Management, Analysis and Visualization		**X**	**X**			**X**	Metodología para abordar proyectos de Big Data de manera sistemática	•Etapas de datos Adquisición y gestión de fuentes de datos • Agregar valor a los datos • Selección e implementación de un Big Data Warehouse •Desarrollo de visualizaciones para Big Data.	Bases de datos NoSQL RDBMS extendido Hadoop / MapReduce.
34	A Storage Model for Handling Big Data Variety			**X**		**X**		Propuesta de modelo de almacenamiento basado en XML para resolver el problema de almacenar cualquier tipo de datos.	Algoritmos de organización y procesamiento de la información en formato XML	Uso de bases de datos no relacionales: Cassandra, SQL, MongoDB, Neo4j, HDFS.

35	Beyond the hype: Big Data concepts, methods, and analytics			**X**	**X**			Describir los métodos analíticos más utilizados para el análisis de Big Data	•Análisis de texto •Análisis de audio •Analítica de video •Analítica predictiva	• Técnicas y algoritmos de extracción de información. • Enfoque basado en la transcripción •Enfoque basado en la fonética
36	Big Data preprocessing: methods and prospects (IMAGS)			**X**	**X**			Mostrar los métodos y técnicas de pre procesamiento de datos para la minería de datos en Big Data.	Pre procesamiento de datos. Transformación de datos, integración, limpieza y normalización.	• TF-IDF • Word2Vec • CountVectorizer • Tokenizer •StopWordsRemover • n-gram

A continuación, se muestra la Gráfica 22 y la Tabla 36 donde se presenta el total de las herramientas, frameworks y metodologías que son utilizados para abordar los diferentes tipos de datos.



Gráfica 22. Total de Herramientas, Frameworks y Metodologías por tipos de datos

Tabla 36. Herramientas, Frameworks y Metodologías por tipo de datos

	Estructurados	**Semiestructurados**	**No estructurados**
Herramientas	Hadoop	Hadoop	Hadoop
	Spark	Spark	Spark
	Genus	Genus	Genus
	IBM InfoSphere	IBM InfoSphere	IBM InfoSphere
	Storm	------	Storm
	IBM BigSheets	IBM BigSheets	------
	------	Project Voldemort	Project Voldemort
	------	Cloudera	------
	-------	-------	Flink
	-------	Samza	-------
	-------	-------	Mahout
	-------	-------	Hortonworks
	-------	-------	MapReduce
	-------	-------	EpiC

Frameworks	Marimba FHDHC FIBD BigDAF Dryad y DryadLINQ Nephele Framework (AaaS) RUBA Twister.Net ----- ----- ----- ----- ----- -----	Marimba FHDHC FIBD BigDAF Dryad y DryadLINQ Nephele Framework (AaaS) -------- -------- FC-BESSSED -------- -------- -------- -------- --------	Marimba FHDHC FIBD BigDAF Dryad y DryadLINQ Nephele Framework (AaaS) RUBA -------- -------- FUDA Framework ETL jMetalSP Piccolo FERIUUBD
Metodologías	AAMBDA BDMPT MDANJHVD ---------- ---------- ---------- ----------	AAMBDA BDMPT MDANJHVD AIMBDMAV ---------- ---------- ----------	AAMBDA BDMPT MDANJHVD AIMBDMAV ASMHBDV BHBDCMA BDPMP(IMAGS)

De acuerdo a la Gráfica 21 y a la Tabla 36 se puede observar lo siguiente:

- Para el tratamiento de los 3 tipos de datos con herramientas, frameworks y metodologías se obtuvieron:

- **4 herramientas** (Hadoop, Spark, Genus y IBM InfoSphere), **7 frameworks** (Marimba, Framework to Handle Data Heterogeneity Contextual "FHDHC", Framework of Integrated Big Data "FIBD", BigDAF, Dryad, Nephele y Framework (AaaS)) y **3 metodologías** (AAMBDA, BDMPT, MDANJHVD).

- Para el tratamiento de 2 tipos de datos con herramientas, frameworks y metodologías se obtuvieron:

- **3 herramientas** (Storm, IBM BigSheets, Project Voldemort), **1 framework** (RUBA) y **1 metodología** (AIMBDMAV).

- Para el tratamiento de 1 tipo de dato con herramientas, frameworks y metodologías se obtuvieron:

- **7 herramientas** (Cloudera, Flink, Samza, Mahout, Hortonworks, MapReduce y EpiC), **7 frameworks** (Twister.Net, Framework for Concept-Based Exploration of Semi-Structured "FC-BESSSED", Framework for Unstructured Data Analysis "FUDA", Framework ETL, jMetalSP, Piccolo, Framework for Extracting Reliable Information from Unstructured Uncertain Big Data "FERIUUBD") y **3 metodologías** (ASMHBDV, BHBDCMA, BDPMP (IMAGS)).

En resumen, la plataforma de código abierto Hadoop tiene el liderazgo en la actualidad como el mejor entorno para analizar grandes cantidades de datos. Ya que, está inspirado en el paradigma de programación MapReduce, el cual consiste en dividir en dos tareas (mapa - reducir) logrando un alto paralelismo en el procesamiento. (M. Khan, 2014)

Los frameworks para procesar los distintos tipos de datos utilizan Hadoop, ya que Hadoop se puede combinar perfectamente con distintas herramientas y plataformas. Por ejemplo, SAS incorpora Hadoop en sus aplicaciones. También SAS permite trabajar en memoria a través de Hadoop. IBM trabaja con Hadoop en su plataforma IBM InfoSphere BigInsights (BigInsights). Microsoft incluye Hadoop en SQL Server 2014, Windows Server 2012, HDInsight and Polybase.

Con los resultados obtenidos se pretende abordar el problema de la variedad de datos en Big Data implementando un grupo de modelos de acuerdo a las características presentadas anteriormente.

Los métodos y modelos más utilizados para Big Data en general son los siguientes:

a) Visualización de los datos: La visualización de los datos se ha convertido en uno de los métodos más aprovechados debido a su facilidad de interpretar y detectar patrones en los datos.

b) Análisis de series temporales: Consiste de un conjunto de técnicas estadísticas que analizan secuencias de datos para predecir la probabilidad de un resultado. La analítica predictiva se establece y aprende del pasado para proyectar determinados escenarios en un futuro, ayudando en la toma de decisiones.

c) Análisis de textos: La minería de textos consta de técnicas de análisis de la información mediante procesos semánticos de grandes volúmenes de textos. Esto con el fin de extraer valor de la información procesada para un determinado fin.

d) Algoritmos de procesamiento: Se trata de algoritmos capaces de detectar una relación entre distintas variables de procesamiento de datos en una variedad de distintas bases de datos. Esto con el fin de estructurar diversos tipos de datos y agruparlos en bases de datos unificados.

6.2 Propuestas para abordar el problema de la variedad

Durante el análisis y desarrollo del mapeo sistemático se encontró una metodología, cuatro métodos y dos modelos para abordar el problema de la variedad, pero no se encontró una metodología estándar o definitiva para abordar el problema de la variedad para los 3 tipos de datos (estructurados, semiestructurados y no estructurados), sin embargo, con la información de los trabajos analizados se pueden utilizar varias herramientas, frameworks, modelos y métodos para tratar de abordar este problema.

El desafío más grande es saber con qué tecnologías se puede analizar la información a partir de datos tan variados, así como también el procesamiento de grandes cantidades de datos para la toma de decisiones y el análisis predictivo. Para ello las características principales de los datos se presentan en 3 tipos: Estructurados, Semiestructurados y No estructurados.

Los datos estructurados se pueden considerar generalmente como archivos de texto, que están bien organizados, los datos no estructurados son incompletos y de naturaleza desorganizada y por último los datos semiestructurados son una mezcla de los dos anteriores y no presentan una estructura perfectamente definida como los datos estructurados, pero si presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones. (M. Suyash, 2017)

Durante el estudio de mapeo sistemático se encontraron herramientas, frameworks, métodos, algoritmos y técnicas, que pueden procesar y proporcionar un análisis y en su mayoría un almacenamiento eficiente de datos no estructurados.

En el contexto anterior se llevó un análisis de los estudios y se encontró que mediante el procesamiento de datos no estructurados para transformarlos a un tipo de dato semiestructurado puede ayudar a la organización y almacenamiento de los datos en un formato de Clave – Valor para su procesamiento y extracción de conocimiento.

La Figura 7 muestra un conjunto de herramientas, frameworks y metodologías para abordar el problema de la variedad por medio de su tipo de datos.

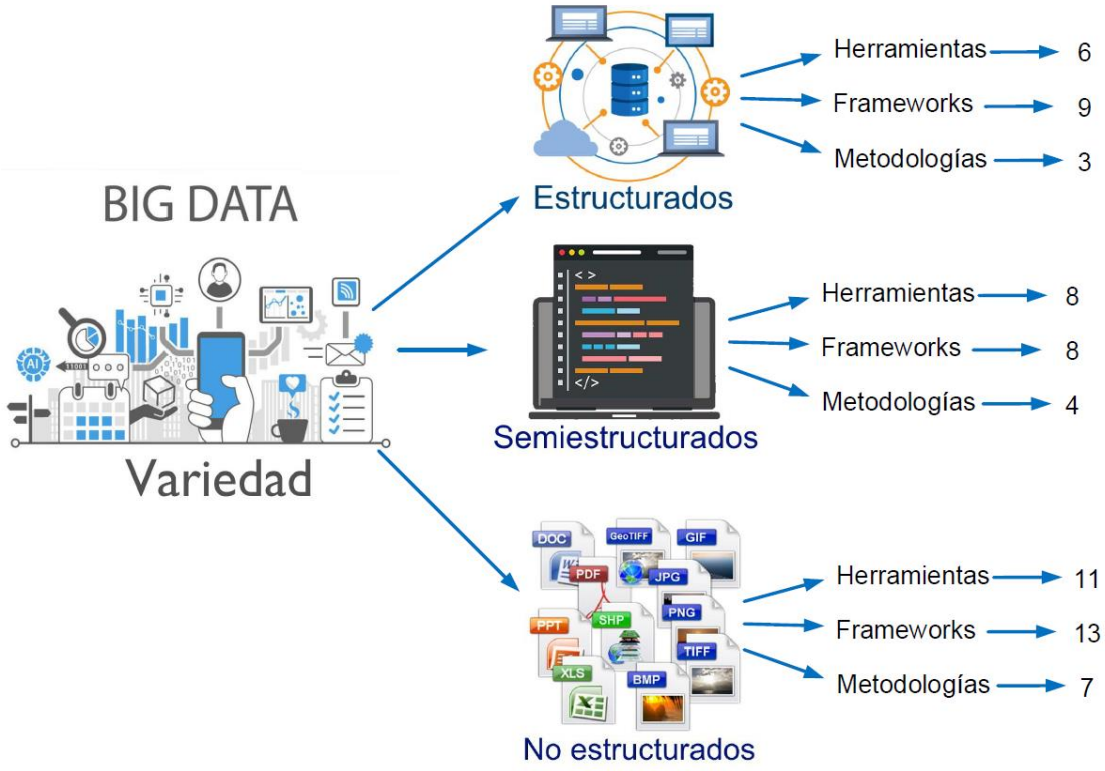


Figura 7. Herramientas, Frameworks y Metodologías para tratar datos Estructurados, Semiestructurados y No estructurados

La fase inicial de este trabajo es la organización de las Herramientas, Frameworks y Metodologías para poder abordar el problema de la variedad desde la perspectiva de los diferentes tipos de datos, cabe mencionar que algunas herramientas y frameworks pueden procesar varios tipos de datos de una forma similar.

La Figura 8 muestra de una forma más detallada los frameworks y herramientas para procesar los tipos de datos estructurados dependiendo del formato que se quiera abordar.

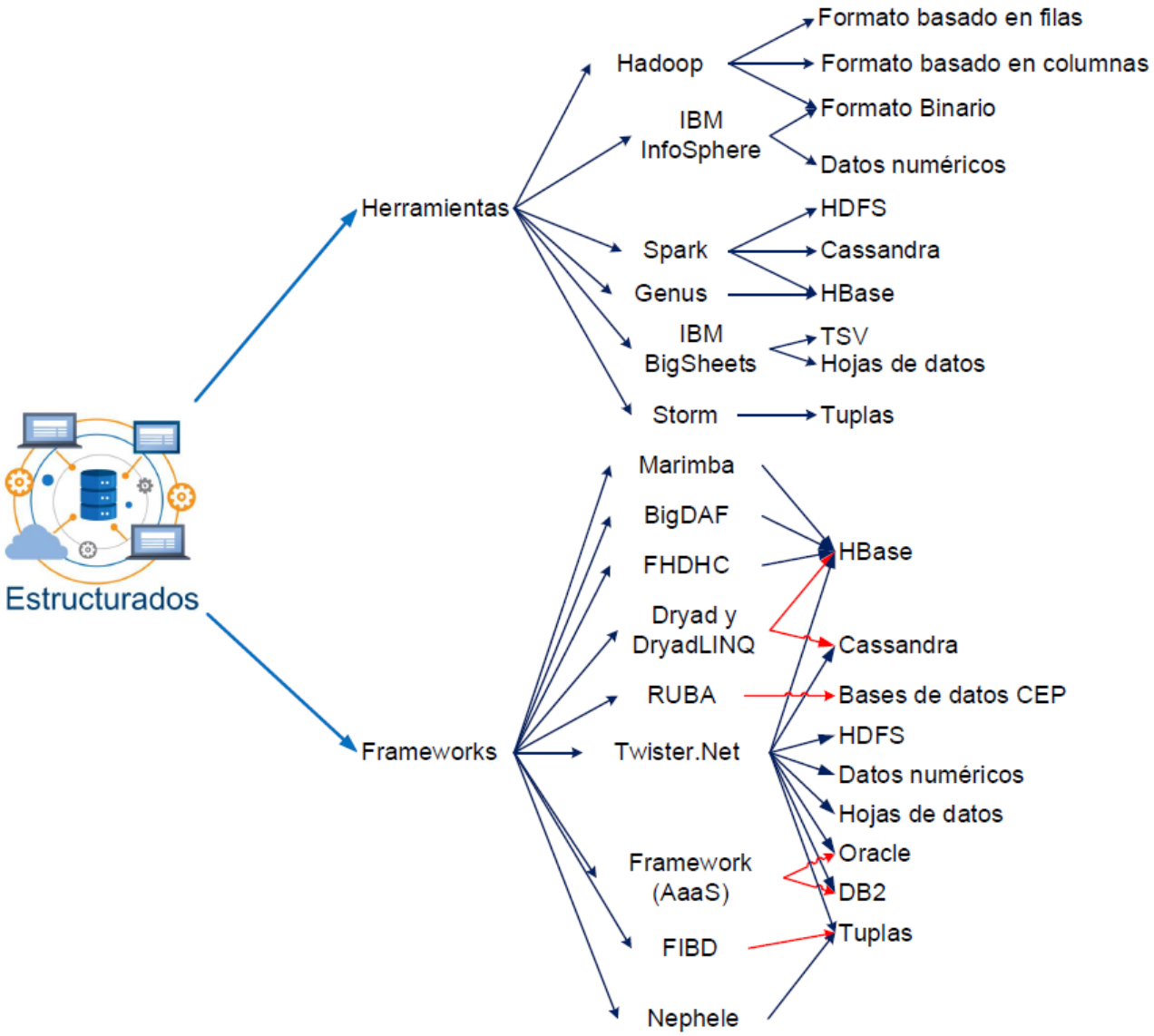


Figura 8. Herramientas y Frameworks para tratar datos Estructurados en diversos formatos.

Para información más detallada sobre el uso de herramientas y frameworks propuestos ver el Anexo A) Descripción para el procesamiento de datos estructurados (ver Anexo).

La Figura 9 muestra de una forma más detallada las herramientas y frameworks para procesar los tipos de datos Semiestructurados.

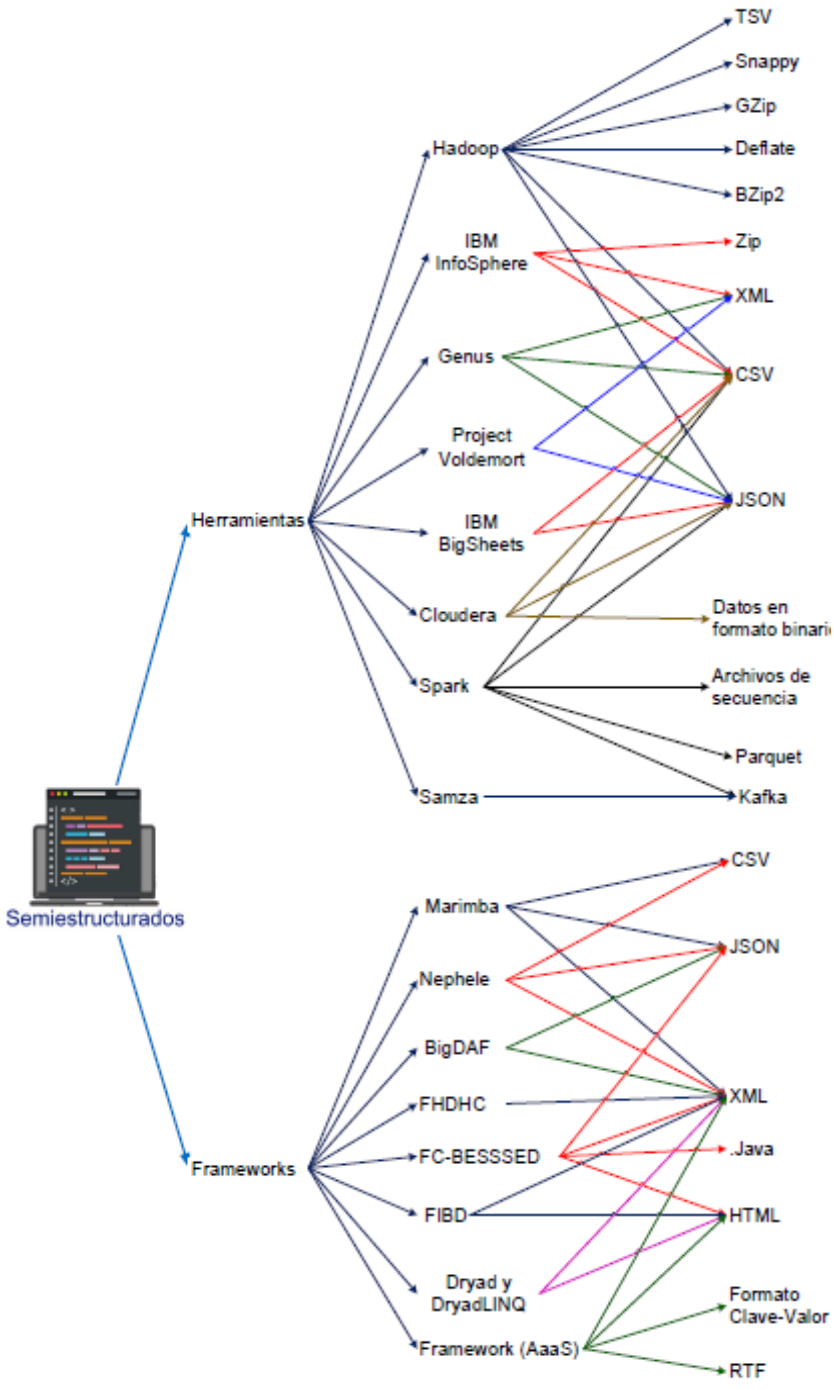


Figura 9. Herramientas y Frameworks para tratar diversos formatos de datos Semiestructurados

Para información más detallada sobre el uso de herramientas y frameworks propuestos ver el Anexo B) Descripción para el procesamiento de datos semiestructurados (ver Anexo).

La Figura 10 muestra de una forma más detallada las herramientas y frameworks para procesar los tipos de datos No estructurados.

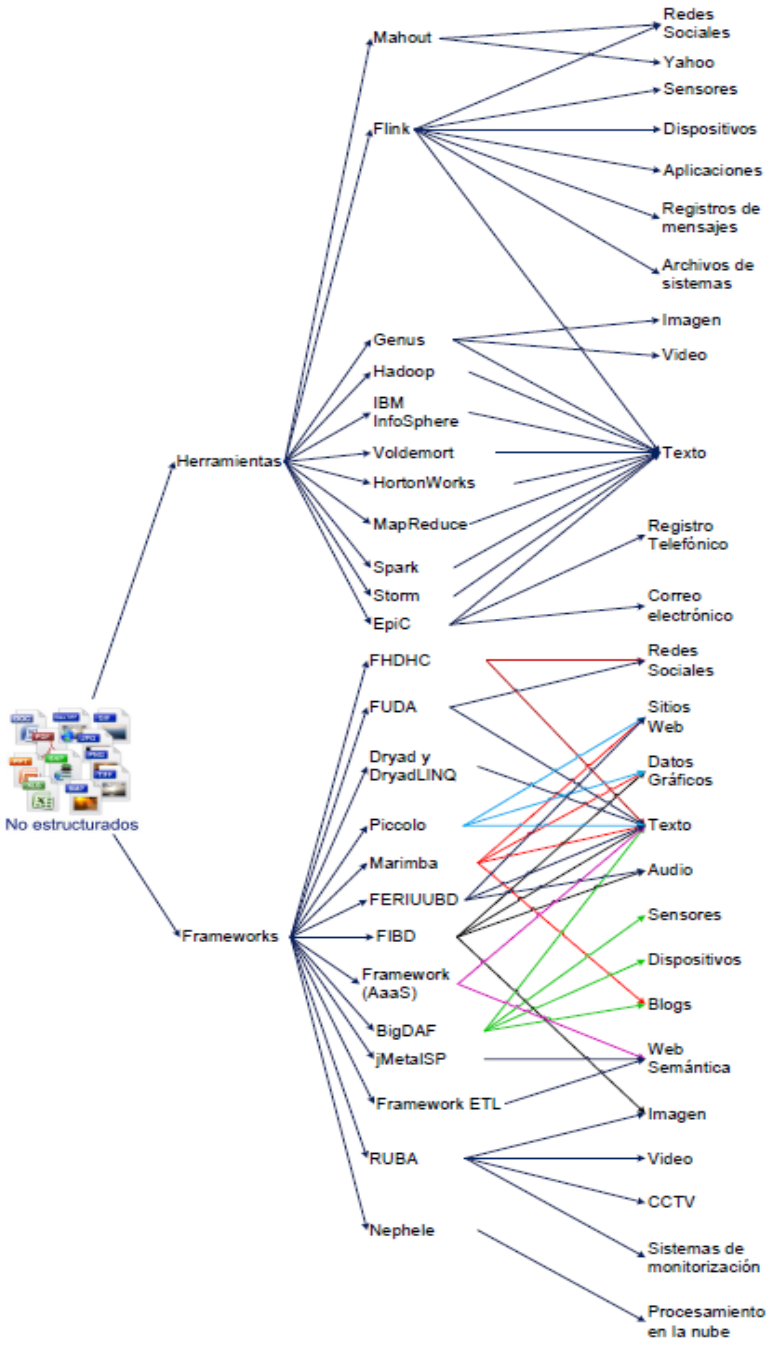


Figura 10. Herramientas y Frameworks para tratar diversas fuentes de datos No-estructurados

Para información más detallada sobre el uso de herramientas y frameworks propuestos ver el Anexo C) Descripción para el procesamiento de datos no estructurados (ver Anexo).

Capítulo 7

Conclusiones

En este capítulo se presentan las conclusiones a las que se llegaron, asi como también a las aportaciones adquiridas como resultado de este trabajo de investigación, además, se sugieren algunos trabajos futuros para dar continuidad a la investigación.

7.1 Conclusiones

De acuerdo con los resultados presentados en el Capítulo 5 se concluye el cumplimiento del objetivo general, que se menciona en la sección 1.3 del Capítulo 1:
“Evidenciar el estatus actual del problema de la variedad de tipos de datos en sistemas Big Data”, “Realizar un estudio de mapeo sistemático sobre el problema de la variedad en sistemas Big Data” y “Enfocar el estudio de mapeo a Herramientas, IDE's, frameworks y metodologías utilizados para tratar el problema de la variedad en sistemas Big Data”.

Durante el proceso de investigación y desarrollo del estudio de mapeo sistemático, se tuvieron los siguientes hallazgos:

1. No se obtuvo algún artículo que tenga como fin una revisión sistemática, algún mapa o mapeo sistemático en el problema de la variedad en Big Data, tampoco se encontró algúna guía que ayudara a la selección de herramientas, frameworks o metodologías para abordar el problema de la variedad.

2. Se llevó a cabo el análisis del mapeo sistemático para evidenciar las herramientas, frameworks y metodologías y de esta manera presentar resultados para abordar el problema de la variedad y que sirva como guía para poder trabajar con los distintos tipos y formatos de datos.

3. Se identificaron 14 herraminetas, 24 frameworks y 27 metodologías que que abordan el problema de la variedad en Big Data desde diferentes perspectivas, desde el análisis de datos en bruto por medio del procesamiento con el formato de clave – valor, hasta el análisis de textos, imágenes, audio y diferentes formatos de datos para su extracción, transformación y carga de nuevos tipos de datos en alguna plataforma para su extracción de conocimiento.

4. Se encontró que la única herramienta que trata las imágenes y videos es Genus, proponiendo que el video se trabaje como una serie de imágenes y transformar los datos obtenidos en formatos XML para poder tener un formato semiestructurado.

5. Con respecto a las metodologías, cada una de ellas realiza su procesamiento mediante algoritmos matemáticos y métodos que en su mayoría tratan de unificar los datos mediante un almacenamiento general para agrupar distintos tipos de datos. Aunque tradicionalmente el almacenamiento de datos forma el primer desafío en la gestión de un gran volumen de datos, así como su procesamiento.

Actualmente el problema de la variedad es un desafío latente en el desarrollo de sistemas Big Data, esto debido a los avances tecnológicos y al crecimiento exponencial de los diferentes tipos de datos, así como también al surgimiento de nuevos tipos de datos llamados datos heterogéneos.

Existen decenas de herramientas y frameworks para procesar los datos de Big Data, pero el problema general para abordar la variedad de datos es saber cuál escoger. En este trabajo se describieron un conjunto de herramientas y Frameworks para abordar el problema de los diferentes tipos de datos y formatos para llevar a cabo un análisis de la información para poder extraer datos relevantes.

En este trabajo se observó que no existe un framework, herramienta o conjunto de herramientas estándar que puedan servir como base única para abordar el problema de la variedad de datos, la diversidad de herramientas y sus alcances son variados.

7.2 Aportaciones

Las aportaciones principales de este trabajo de tesis son:
1. Un estudio de mapeo sistemático en forma de informe visual que sirve para evidenciar un conjunto de herramientas, frameworks y metodologías que en la actualidad son utilizados para el análisis y procesamiento de grandes conjuntos de datos.
2. Una clasificación de estudios donde se encontraron un conjunto de herramientas, framework y metodologías para el tratamiento, análisis y procesameinto de los diferentes tipos de datos.
3. Otra contribución importante además del mapeo sistemático es la propuesta de un conjunto de herramientas, frameworks y metodologías que sirven para abordar el problema de la variedad por medio de su tipo de datos, que a su vez sirva una guía inicial del panorama del problema de la variedad en Big Data.

7.3 Trabajos Futuros

Para los trabajos futuros se propone:

1. Detallar y evaluar las propuestas sobre las herramientas, frameworks y metodologías que sirven para abordar el problema de la variedad.
2. Utilizar el conjunto de herramientas y frameworks propuestos para estructurar la mayoría de los tipos de datos mencionados, y a su vez comprobar que es posible tratar los videos como imágenes secuenciales con ayuda de la herramienta Genus, ya que este formato es el que existe en mayor cantidad y el menos abordado con alguna herramienta o algún framework.
3. Desarrollar trabajo de investigación acerca de un almacen (Base de datos) que pueda agrupar los diferentes tipos de datos mediante diferentes modelos de almacenamientos como lo realiza el multimodelo de bases de datos (L. Jiaheng, 2019) y el modelo de almacenamiento (S. Anindita Sarkar, 2017), utilizando modelos relacionales basados en términos matemáticos de relaciones (subconjuntos) utilizando un DBMS relacional para asegurar el almacenamiento y la recuperación de los datos mediante modelos semiestructurados para documentos XML y JSON, Modelo de clave / valor y el Modelo de grafos dedicada al almacenamiento y la gestión eficiente de los datos.

Referencias

A. Barbara, B. D. (2010). The value of mapping studies: a participantobserver case study. *ResearchGate*, 25-33.

A. Jacky, C. I. (2017). Research on Big Data - A systematic mapping study. *ElSevier*, 105-115.

A. Nuha, N. A. (2016). A Systematic Mapping Study in Microservice Architecture. *IEEE explore*, 44-51.

A. W. Matthew, E. F. (2013). Data Science, Predictive Analytics, and Big Data: A RevolutionThat Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, 77-84.

B. Elizabeth. Sanders, B. E. (2010). A framework for organizing the tools and techniques of participatory design. *ACM Digital Library*, 195-198.

C. Eaton, D. D. (2012). Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. *Mc Graw-Hill Companies*, 978-985.

C. Manterola, P. A. (2013). Revisiones sistemáticas de la literatura. Qué se debe saber acerca de ellas. *ElSevier*, 149-155.

C. Min, M. S. (2014). Big Data: A Survey. *SpringerLink*, 171-210.

Caracheo, F. (2002). Modelo educativo (propuesta de diseño). *CIDET*.

Cukier, K. (27 de Febrero de 2010). *Data everywhere: A special report on managing information.* Obtenido de https://www.economist.com/special-report/2010/02/27/data-data-everywhere

D. Arantxa, O. A. (2014). A Big Data methodology for categorising technical support requests using Hadoop and Mahout. *Journal of Big Data a SpringerOpen Journal*, 1-11.

D. Gough, J. T. (2012). Clarifying differences between review designs and methods. *Systematic Reviews*, 1-9.

D. Jeff, G. S. (2004). Parallel execution. In: MapReduce: simplified data processing on large clusters. *Google, Inc*, 1-13.

D. Rustem, D. S. (2017). Quantifying Volume, Velocity, and Variety to Support (Big) Data-Intensive Application Development. *IEEE explore*, 1-10.

Doug, L. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Apication Delivery Strategies ,Meta Group*, 1-4.

E. Manuel, I. M. (2004). Generalidades sobre Metodología de la Investigación. *Universidad Autónoma del Carmen Ciudad del Carmen*, 1-105.

Esther, M. (2014). Métodos y técnicas de investigación. *Universidad Nacional Autónoma de México Facultad de ArquitecturaMéxico*.

F. T. Luis, F. H. (2015). *Descubrimiento de Conocimiento en Big Data: Estudio de Mapeo Sistémico.* Tesis de Maestría, Universidad de San Buenaventura Cali, Cali, Colombia.

Felipe, S. (2015). Big Data. *Dialnet Redexis Gas*, 71-86.

Francisco, G. P. (2017). Revisión sistemática de literatura en los Trabajos de Final de Máster y en las Tesis Doctorales. *Grupo GRIAL, Universidad de Slamanca*, 1-95.

G. Amir Gandomi, H. M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *ElSevier Ltd.*, 137-144.

G. Kim, S. T. (2014). Big-Data applications in the Government sector. *ACM Digital Library*, 78-85.

G. Palak, T. N. (2015). An Approach Towards Big Data-A Review. *IEEE explore*, 118-123.

Gartner, R. (15 de 08 de 2008). *Gartner Research*. Obtenido de https://www.gartner.com/en/information-technology/glossary/big-data

Gerard, J. (2014). Big Data and the SP Theory of Intelligence. *IEEE explore*, 301-315.

Gillian, J. (2015). A Generic Framework for Concept-Based Exploration of Semi-Structured Software Engineering Data. *IEEE explore*, 894-897.

H. Hu, Y. W. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE explore*, 652-687.

H. Lamyae, B. H. (2016). Big Data: framework and issues. *IEEE explore*, 485-490.

H. Muhammad, S. C. (2016). Big Data Reduction Methods: A Survey. *Springer Nature*, 1-20.

Hua, T. C. (2006). Taxonomy of Java web application frameworks. *IEEE explore*, 377-385.

K. Barbara Kitchenham, C. S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *ResearchGate*, 1-44.

K. Manjit, S. (2013). Big Data and Methodology-A review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 991-995.

K. Navroop, K. S. (2019). Cloud resourcemanagement using 3Vs of Internet of Big Data streams. *Springer Nature*, 1463–1485.

K. Petersen, F. R. (2008). Systematic Mapping Studies in Software Engineering. *ResearchGate*, 1-10.

K. Petersen, S. V. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *El Sevier*, 1-18.

K. Ravindra, A. K. (2015). Efforts toward Research and Development on Inconsistencies and Analytical tools of Big Data. *ResearchGate*, 845-850.

K. Stephen, A. F. (2013). Big Data: Issues and Challenges Moving Forward. *IEEE explore*, 995-1004.

Keith, G. (2013). What is Big Data? *ResearchGate*, 12-13.

L. Jiaheng, H. I. (2019). Multi-model Databases: A New Journey to Handle the Variety of Data. *ACM Computing Surveys*, 1-38.

Luis, O. G. (22 de Septiembre de 2020). *Conogasi.org*. Obtenido de http://conogasi.org/articulos/algoritmo/

M. Ahmad, S. J. (2018). ig Data: Issues, Challenges, and Techniques in Business Intelligence. *Springer Nature Singapore*, 613-628.

M. Khan, F. M. (2014). Seven V's of Big Data Understanding Big Data to extract Value. *IEEE explore*, 1-5.

M. Suyash, A. M. (2017). Structured and Unstructured Big Data Analytics. *International Conference on Current Trends in Computer, Electrical, Electronics and Communication*, 1-7.

M. Wolfgang, G. G. (2018). Variety Management for Big Data, Springer Verlag GmbH Germany. *Springer Nature*, 47-62.

Manuel, G. (2008). *Definicion De Terminos Basicos De Investigacion (Glosario)*. Obtenido de EcuRed.

Manuel, G. (2013). *Gestión Científica Empresarial*, 75.

Marketer, B. D. (01 de October de 2018). *Big Data Social*. Obtenido de http://www.bigdata-social.com/

N. Kaur, K. S. (2019). Cloud resourcemanagement using 3Vs of Internet of Big Data streams. *Springer Nature*, 1-23.

P. Shantanu, D. R. (2018). Adaptive System for Handling Variety in Big Text. *Springer Nature Singapore*, 305-313.

R. Lisbeth, R. C. (2015). general perspective of Big Data: applications, tools, challenges and trends. *SpringerLink*, 1-41.

RedHat. (10 de 2020). *RedHat*. Obtenido de https://www.redhat.com/es/topics/middleware/what-is-ide

Rohit, S. (2018). Data Mining Techniques: Types of Data, Methods, Applications. *ElSevier Ltd.*, 1-24.

S. Anindita Sarkar, C. S. (2017). A Storage Model for Handling Big Data Variety. *Springer Nature Singapore*, 59-71.

S. Kamakhya, K. R. (2019). Big Data Ecosystem: Review on Architectural Evolution: Proceedings of IEMIS 2018. *ResearchGate*, 335-345.

S. Onur, S. Y. (2013). Tactical Big Data Analytics: Challenges, Use Cases, and Solutions. *ACM Digital Library*, 1-4.

S. Rolf, S. S. (2009). Combining Unstructured, Fully Structured andSemi-Structured Information in Semantic Wikis. *ResearchGate*, 1-12.

Sáenz, A. (2018). Mapping y revisiones sistemáticas. *Boletín de la Sociedad de Pediatría de Asturias*, 215-221.

Santiago, H. G. (2019). *Mapeo sistemático del reconocimiento del habla, proceso del lenguaje natural y uso de ontologías para identificar el dominio del problema y los requerimientos de solución.* Tesis de Maestría, Centro Nacional de Investigación, Ciencias Computacionales, Cuernavaca, Morelos.

Sebastian, S. L. (2016). Metodología para el modelamiento de datos basado en Big Data, enfocados al consumo de tráfico (voz-datos) generado por los clientes. *Springerlink*, 1-17.

Singh, S. S. (2012). Big Data Analytics. *IEEE explore*, 1-15.

Soha, S. (2016). A Comparative Study to Classify Big Data Using Fuzzy Techniques. *IEEE explore*, 1-6.

Subhash, K. (2016). Evolution of Spark Framework for simplifying Big Data Analytics. *IEEE explore*, 1-6.

Suresh, J. (2014). Bird´s Eye View On "Big Data Management". *IEEE explore*, 1-5.

UNAM. (2001). *Dirección General de Servicios de Cómputo Académico*.

Vicente, T. (2015). Algoritmos. *Autores científico-técnicos y académicos, ACTA*, 43-50.

Victoria, B. (Febrero de 2009). *Definición ABC*. Obtenido de https://www.definicionabc.com/tecnologia/cluster.php

W. A. Mudasir, J. S. (2018). Big Data: Issues, Challenges, and Techniques in Business Intelligence. *Springer Nature Ltd.*, 613-628.

W. Xiaokang, T. L. (2018). A Big Data-as-a-Service Framework: State-of-the-Art and Perspectives. *IEEE explore*, 1-16.

Wilbur, L. (2019). NIST Big Data Interoperability Framework: Volume 4, Security and Privacy. *NIST Special Publication* , 1-176.

X. Chunli, G. J. (2017). Big Data Validation Case Study. *IEEE explore*, 1-6.

Y. Piao, W. T. (2015). From Text to XML by Structural Information Extraction. *IEEE explore*, 448-452.

Yin, R. K. (2012). Case Study Research: Design and Methods, 3/e. Thousand Oaks.
JSTOR, 93-95.

[66] Rachita Misra, Bijayalaxmi Panda, Mayank Tiwary, “Big Data and ICT applications – A study”, ACM Digital Library, pp. 1-6, March 2016.

[67] Dawei Jiang, Sai Wu, Gang Chen, Beng Chin Ooi, Kian-Lee Tan, Jun Xu, “epiC: an extensible and scalable system for processing Big Data”, Springer-Verlag Berlin Heidelberg, pp. 1-24, July 2015.

[68] Fatima Riaz, Muhammad Alam, Attra Ali, “Filtering the Big Data Based on Volume, Variety and Velocity by Using Kalman Filter Recursive Approach”, IEEE explore, pp. 1-6, 2017.

[69] Salwa Souissi, Mounir BenAyed, “GENUS: an ETL tool treating the Big Data Variety”, ACM Digital Library, pp.1-8,2016.

[70] Rustem Dautov, Salvatore Distefano, “Quantifying Volume, Velocity, and Variety to Support (Big) Data-Intensive Application Development”, IEEE International Conference on Big Data (BIGDATA), pp. 1-10, 2017.

[71] Parth Chandarana, M. Vijayalakshmi, “Big Data Analytics Frameworks”, International Conference on Circuits, Systems, Communication and Information Technology Aplications (CSCITA), pp.1-5, 2014.

[72] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, “Big Data and Advanced Analytics Tools”, Symposium on Colossal Data Analysis and Networking (CDAN), SpringerLink, pp. 1-8, 2016.

[73] Wissem Inoubli, Sabeur Aridhi, Haithem Mezni, Mondher Maddouri, Engelbert Mephu Nguifo, “An experimental survey on Big Data frameworks”, Elsevier B.V., pp.1-19, 2017.

[74] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih,"Big Data technologies: A survey", Journal of King Saud University – Computer and Information Sciences 30, ACM Digital Library, pp. 431-448, 2016.

[75] Sarpreet Kaur, Dr. Williamjeet Singh, "Systematic mapping study of Big Data mining tools and techniques", IEEE explore, pp. 1-7, 2017.

[76] Jaskaran Singh, Varun Singla, "Big Data: Tools and Technologies in Big Data", International Journal of Computer Applications (0975 – 8887), Volume 112 – No 15, pp. 1-5, February 2015.

[77] Ciprian Dobre, Fatos Xhafa, "Parallel Programming Paradigms and Frameworks in Big Data Era", Springer Science+Business Media, pp 1-29, New York 2013.

[78] T. Ramalingeswara Rao, Pabitra Mitra, Ravindara Bhatt, A. Goswami, "The Big Data system, components, tools, and technologies: a survey", Springer-Verlag London Ltd., part of Springer Nature, pp. 1-81, 2018.

[79] Mahidhar Tatineni, Xiaoyi Lu, Dongju Choi, Amit Majumdar, Dhabaleswar K. (DK) Panda, "Experiences and Benefits of Running RDMA Hadoop and Spark on SDSC Comet", ACM Digital Library, pp. 1-5, July 2016.

[80] Rui Mao, Honglong Xu, Wenbo Wu, Jianqiang Li, Yan Li, and Minhua Lu, "Overcoming the Challenge of Variety: Big Data Abstraction, the Next Evolution of Data Management for AAL Communication Systems", IEEE Communications Magazine, pp. 1-6, January 2015.

[81] Sourav Mazumder, "Big Data Tools and Platforms", Springer International Publishing Switzerland, pp. 1-100, 2016.

[82] Guoliang CHEN, RuiMAO, Kezhong LU, "A parallel computing framework for Big Data", Higher Education Press and Springer-Verlag Berlin Heidelberg, pp. 1-14, 2016.

[83] Ravindra Kumar Yadav, Khan Atiya Naaz, "Efforts toward Research and Development on Inconsistencies and Analytical tools of Big Data", International Conference on Computational Intelligence and Communication Networks, pp. 1-6, 2015.

[84] Tanmaya Mahapatra, Ilias Gerostathopoulos, Christian Prehofer, "Towards Integration of Big Data Analytics in Internet of Things Mashup Tools", ACM Digital Library, pp.1-6, 2016.

[85] Thomas Cerqueus, Eduardo Cunha de Almeida, and Stefanie Scherzinger, "Safely Managing Data Variety in Big Data Software Development", IEEE/ACM 1st International Workshop on Big Data Software Engineering, pp. 1-7, 2015.

[86] Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter, and Tawfiq Hasanin, "A survey of open source tools for machine learning with Big Data in the Hadoop ecosystem", Journal of Big Data a SpringerOpen Journal, pp. 1-36, 2015.

[87] Gurjit Singh Bhathal, and Amardeep Singh, "Big Data Computing with Distributed Computing Frameworks", Springer Nature Singapore, pp. 1-11, 2019.

[88] Lamyae HBIBI, Hafid BARKA, "Big Data: framework and issues", 2nd International Conference on Electrical and Information Technologies ICEIT'2016, JSTOR, pp. 1-6, 2016.

[89] Gurjit Singh Bhathal, Amardeep Singh, "Big Data: Hadoop framework vulnerabilities, security issues and attacks", Elsevier B.V, pp. 1-8, 2019.

[90] Chitresh Verma, Dr. Rajiv Pandey, "Big Data representation for Grade Analysis Through Hadoop Framework", JSTOR, pp. 1-4, 2016.

[91] Venketesh Palanisamy, Ramkumar Thirunavukarasu, "Implications of Big Data analytics in developing healthcare frameworks – A review", Journal of King Saud University – Computer and Information Sciences 31, ACM Digital Library, pp. 415-125, 2017.

[92] Yunqing Liu, Jianhua Zhang, Shuqing Han, Mengshuai Zhu, "Research on the Computing Framework in Big Data Environment", 3rd International Conference on Information Science and Control Engineering, pp. 1-5, 2016.

[93] Jeffrey S. Saltz, "The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness", IEEE International Conference on Big Data (Big Data), pp. 1-6, 2015.

[94] Lisbeth Rodríguez-Mazahua, Cristian-Aarón Rodríguez-Enríquez, José Luis Sánchez-Cervantes, Jair Cervantes, Jorge Luis García-Alcaraz, Giner Alor-Hernández, "A general perspective of Big Data: applications, tools, challenges and trends", Springer Science+Business Media New York, pp. 1-41, 2015.

[95] Gillian J. Greene, "A Generic Framework for Concept-Based Exploration of Semi-Structured Software Engineering Data", 30th IEEE/ACM International Conference on Automated Software Engineering, pp. 1-4, 2015.

[96] Seung Ryul Jeong, Imran Ghani, "Semantic Computing for Big Data: Approaches, Tools, and Emerging Directions (2011-2014)", KSII Transactions on internet and information systems vol. 8, no. 6, pp. 1-21, 2014.

[97] Subhash Kumar, "Evolution of Spark Framework for simplifying Big Data Analytics", JSTOR, pp. 1-6, 2016.

[98] Onur Savas, Yalin Sagduyu, Julia Deng, and Jason Li, "Tactical Big Data Analytics: Challenges, Use Cases, and Solutions", ACM Digital Library, pp. 1-4, 2013.

[99] T. Ramalingeswara Rao, Pabitra Mitra, Ravindara Bhatt, A. Goswami, "The Big Data system, components, tools, and technologies: a survey", Springer-Verlag London Ltd., part of Springer Nature, pp. 1-81, 2018.

[100] Chunli Xie, Jerry Gao, Chuanqi Tao, "Big Data Validation Case Study", IEEE Third International Conference on Big Data Computing Service and Applications, pp. 1-6, 2017.

[101] Shraddha Dwivedi, Paridhi Kasliwal, Prof Suryakant Soni, "Comprehensive Study of Data Analytics Tools (RapidMiner, Weka, R tool, Knime)", Symposium on Colossal Data Analysis and Networking (CDAN), JSTOR, pp. 1-8, 2016.

[102] Jacky Akoka, Isabelle Comyn-Wattiau, Nabil Laoufi, "Research on Big Data – A systematic mapping study", Elsevier B.V., pp. 1-11, 2017.

[103] Johannes Schildgen, Thomas Jörg, Manuel Hoffmann, Stefan Deßloch, "Marimba: A Framework for Making MapReduce Jobs Incremental", IEEE International Congress on Big Data, pp. 128-135, 2014.

[104] T.K Das, P. Mohan Kumar, "BIG Data Analytics: A Framework for Unstructured Data Analysis", International Journal of Engineering and Technology (IJET), pp. 153-156, February 2013.

[105] Richard K. Lomotey, Ralph Deters, "Analytics-as-a-Service (AaaS) Tool for Unstructured Data Mining", IEEE International Conference on Cloud Engineering, pp. 319-324, 2014.

[106] J Jaein Kim, Nacwoo Kim, Joonho Park, Kwangik Seo, Hunyoung Park, "RUBA: Real-time Unstructured Big Data Analysis Framework", IEEE explore, pp. 518-522, 2013.

[107] Gillian J. Greene, "A Generic Framework for Concept-Based Exploration of Semi-Structured Software Engineering Data", International Conference on Automated Software Engineering, pp. 894-897, 2015.

[108] Prof. Sindhu. C.S, Dr. Nagaratna P. Hegde, "A Framework to Handle Data Heterogeneity Contextual to Medical Big Data", IEEE explore, pp. 1-7, 2015.

[109] Srividya K Bansal, "Towards a Semantic Extract-Transform-Load (ETL) framework for Big Data", ElSevier, pp. 522-529, 2013.

[110] Zhikui Chen, Fangming Zhong, Xu Yuan, Yueming Hu, "Framework of Integrated Big Data: A Review", ACM Digital Library, pp. 1-5, 2014.

[111] Cristóbal Barba González, José Francisco Aldana Montes, José Manuel Gracía Nieto, "Big Data Optimization: Algorítmico para el análisis de Datos guiado por Semántica", ACM Digital Library, pp. 1-6, March 2015.

[112] Filipe Portela, Luciana Lima, Manuel Filipe Santos, "Why Big Data? Towards a project assessment framework", Elsevier B.V., pp. 604-609, 2016.

[113] Ciprian Dobre, Fatos Xhafa, "Parallel Programming Paradigms and Frameworks in Big Data Era", Springer Science+Business Media New York 2013, pp. 1-23, July 2013.

[114] Sanjay Kumar Singh, Neel Mani, Bharat Singh, "A Framework for Extracting Reliable Information from Unstructured Uncertain Big Data", Springer International Publishing Switzerland 2016, pp. 175-185, 2016.

[115] Wissem Inoubli, Sabeur Aridhi, Haithem Mezni, Mondher Maddouri, Engelbert Mephu Nguifo, "An experimental survey on Big Data frameworks", Elsevier B.V., pp.546-564, 2018.

[116] Soha Safwat Labib, "A Comparative Study to Classify Big Data Using Fuzzy Techniques", IEEE explore, pp. 1-6, 2016.

[117] Xiaokang Wang, Laurence T. Yang, Huazhong Liu, and M. Jamal Deen, "A Big Data-as-a-Service Framework: State-of-the-Art and Perspectives", ACM Digital Library, pp. 1-16, 2017.

[118] Rustem Dautov, Salvatore Distefano, "Quantifying Volume, Velocity, and Variety to Support (Big) Data-Intensive Application Development", IEEE International Conference on Big Data (BIGDATA), pp. 1-10, 2017.

[119] Ananta Chandra Das, Sachi Nandan Mohanty, Arupananda Girish Prasad, Aparimita Swain, "A Model for Detecting and Managing Unrecognized Data in a Big Data framework", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), ACM Digital Library, pp. 1-6, 2016.

[120] Axel Oehmichen, Florian Guitton, Kai Sun, Jean Grizet, Thomas Heinis, Yike Guo, "eTRIKS Analytical Environment: A Modular High Performance Framework for Medical Data Analysis", ACM International Conference on Big Data (BIGDATA), pp. 1-8, 2017.

[121] Verena Kantere, and Maxim Filatov, "A Framework for Big Data Analytics", ACM Digital Library, pp. 1-8, 2015.

[122] Thuan L Nguyen, "A Framework for Five Big V's of Big Data and Organizational Culture in Firms", ACM Conference on Big Data (Big Data), pp. 1-3, 2018.

[123] Guoliang CHEN, RuiMAO, Kezhong LU, "A parallel computing framework for Big Data", Higher Education Press and Springer-Verlag Berlin Heidelberg, pp. 1-14, 2016.

[124] T.K. Das, P. Mohan Kumar, "BIG Data Analytics: A Framework for Unstructured Data Analysis", International Journal of Engineering and Technology (IJET), pp. 153-156, February 2013.

[125] Gurjit Singh Bhathal, and Amardeep Singh, "Big Data Computing with Distributed Computing Frameworks", Springer Nature Singapore, pp. 1-11, 2019.

[126] Tobias M. Scholz, "Theoretical Framework", JSTOR, pp. 1-75, 2019.

[127] Bogdan Ionescu, Dan Ionescu, Cristian Gadea, Bogdan Solomon, Mircea Trifan, "An Architecture and Methods for Big Data Analysis", Springer International Publishing Switzerland 2016, pp. 491-514, 2016.

[128] Bogdan Ionescu, Dan Ionescu, Cristian Gadea, Bogdan Solomon, Mircea Trifan, "Big Data: Methods, Prospects, Techniques", Springer International Publishing AG, part of Springer Nature 2018, pp. 305-312, 2018.

[129] Anindita Sarkar, Samiran Chattopadhyay, "A Storage Model for Handling Big Data Variety", Springer Nature Singapore Pte Ltd. 2017, pp. 59-71, 2017.

[130] Jiaheng Lu, Irena Holubová, "Multi-model Databases: A New Journey to Handle the Variety of Data", ACM Computing Surveys, pp. 1-38, June 2019.

[131] Amir Gandomi, Murtaza Haider, "Beyond the hype: Big Data concepts, methods, and analytics", Elsevier B.V., pp. 137-144, 2015.

[132] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, Francisco Herrera, "Big Data preprocessing: methods and prospects", Elsevier B.V., pp. 1-22, 2016.

[133] Roberto Tardío, Alejandro Maté, Juan Trujillo, "An Iterative Methodology for Big Data Management, Analysis and Visualization", IEEE International Conference on Big Data (Big Data), pp. 545-550, 2015.

[134] Arantxa Duque Barrachina, and Aisling O'Driscoll, "A Big Data methodology for categorising technical support requests using Hadoop and Mahout", Journal of Big Data a SpringerOpen Journal, pp. 1-11, 2014.

[135] Soha Safwat Labib, "A Comparative Study to Classify Big Data Using Fuzzy Techniques", IEEE explore, pp. 1-6, 2016.

[136] Ananta Chandra Das, Sachi Nandan Mohanty, Arupananda Girish Prasad, Aparimita Swain, "A Model for Detecting and Managing Unrecognized Data in a Big Data framework", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 1-6, 2016.

[137] Somnath Mazumdar, Daniel Seybold, Kyriakos Kritikos, and Yiannis Verginadis, "A survey on data storage and placement methodologies for Cloud-Big Data ecosystem", Journal of Big Data a SpringerOpen Journal, pp. 1-37, 2019.

[138] Shantanu Pathak, D. Rajeshwar Rao, "Adaptive System for Handling Variety in Big Text", Springer Nature Singapore, pp 1-9, 2018.

[139] Hong-Mei Chen, Rick Kazman, "Agile Big Data Analytics for Web-Based Systems: An Architecture-Centric Approach", IEEE explore, pp. 1-15, 2016.

[140] Bogdan Ionescu, Dan Ionescu, Cristian Gadea, Bogdan Solomon, and Mircea Trifan, "An Architecture and Methods for Big Data Analysis", Springer International Publishing Switzerland, pp. 1-24, 2016.

[141] Muhammad Habib ur Rehman, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah, Samee U. Khan, "Big Data Reduction Methods: A Survey", Springer Nature, pp. 1-20, 2016.

[142] Navroop Kaur, Sandeep K. Sood, Prabal Verma, "Cloud resourcemanagement using 3Vs of Internet of Big Data streams", Springer-Verlag GmbH Austria, part of Springer Nature, pp. 1-23, 2013.

[143] Gokul Parambalath, E. Mahesh, P. Balasubramanian, and P. N. Kumar, "Big Data Analytics: A Trading Strategy of NSE Stocks Using Bollinger Bands Analysis", ACM Digital Library, pp. 155-176, 2013.

[144] Jiazhao Li, Jiuxia Zhao, Minjia Mao, Xi Zhao, Jianhua Zou, "Exploring the Prediction o f Variety Variety-seeking Behavior", ACM Digital Library, pp. 1-5, 2019.

[145] Marcos Aurélio Almeida da Silva, Andrey Sadovykh, Alessandra Bagnato, Alexey Cheptsov, Ludwig Adam, "JUNIPER: Towards Modeling Approach Enabling Efficient Platform for Heterogeneous Big Data Analysis", ACM Digital Library, pp. 1-7, 2016.

[146] Qiumei Ouyang, Chao Wu, Lang Huang, "Methodologies, principles and prospects of applying Big Data in safety science research ACM Digital Library, pp. 1-12, 2017.

[147] M. R. Martinez-Torres, D. G. Reina, S. L. Toral, F. Barrero, "Metodologías de Análisis de los Big Data en las Plataformas Educativas", ACM Digital Library, pp. 1-5, 2014.

[148] Alex Endert, Samantha Szymczak, Dave Gunning, John Gersh, "Modeling in Big Data Environments", ACM Digital Library, pp. 1-3, 2016.

[149] Jae-Hoon An, Jae-Gi Son, Ji-Woo Kang, "Squall: Stream Processing and Analysis Model Design", ACM Digital Library, pp. 1-3, 2017.

[150] Suyash Mishra, Dr Anuranjan Misra, "Structured and Unstructured Big Data Analytics", International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC-2017), ACM Digital Library, pp. 1-7, 2017.

[151] Qingquan Song, Hancheng Ge, James Caverlee, and Xia Hu, "Tensor Completion Algorithms in Big Data Analytics", ACM Transactions on Knowledge Discovery from Data, Vol. 13, No. 1, Article 6, pp. 1-48, 2019.

[152] James de Castro Martins, Adriano Fonseca Mancilha Pinto, Edizon Eduardo Basseto Junior, Gildarcio Sousa Goncalves, Henrique Duarte Borges Louro, Jose Marcos Gomes, Lineu Alves Lima Filho, Luiz Henrique Ribeiro Coura da Silva, Romulo Alceu Rodrigues, Wilson Cristoni Neto, Adilson Marques da Cunha, and Luiz Alberto Vieira Dias, "Using Big Data, Internet of Things, and Agile for Crises Management", Springer International Publishing AG 2018, pp. 373-382, 2018.

[153] Jeffrey S. Saltz, "The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness", IEEE International Conference on Big Data (Big Data), pp. 1-6, 2015.

Anexos

Anexo A) Descripción para el procesamiento de datos estructurados

A continuación, se se muestran las herramientas y frameworks propuestos para procesar los datos estructurados de acuerdo al conjunto propuesto para abordar el problema de la variedad (Capitulo 6) y a la Figura 8.

Para el procesamiento de datos estructurados se obtuvieron **6 herramientas, 9 frameworks y 3 metodologías** las cuales son:

a) Herramientas:

1. Hadoop: el formato de entrada pueden ser archivos en formato binario, formato basado en filas, basado en columnas, tablas, bases de datos distribuidas, bases de datos relacionales, bases de datos no relacionales, etc., el procesamiento es mediante la división de los datos mediante clúster utilizando MapReduce para separar la información en un formato clave – valor y de esta manera almacenar la información en un formato binario dentro de un almacén de datos con formato Clave - Valor.

2. Spark: los formatos de entrada pueden ser obtenidas desde registros de HDFS (Hadoop Distributed File System), Bases de datos HBase y Cassandra, el procesamiento es mediante Hadoop y MapReduce para separar la información en un formato CSV (Comma-Separated Values) para representar datos en forma de tabla y en un conjunto de datos distribuidos resilientes(RDD) para distribuir la información, modificarla y obtener alguna salida de ella.

3. Genus: el formato de entrada pueden ser datos obtenidos desde Bases de datos HBase, el procesamiento es mediante algoritmos de Extracción Transformación y carga para separar la información dando como resultado un formato XML.

4. IBM InfoSphere: La entrada de datos puede ser mediante 2 tipos: Lotes y en tiempo real, el formato de entrada es mediante formato binario y formatos de datos numéricos, el

procesamiento es mediante la extracción y transformación de los datos mediante un lenguaje de consulta declarativa para entregar un formato binario o CSV.

5. Storm: los formatos de entrada pueden ser obtenidas desde tuplas, el procesamiento es con ayuda de Apache ZooKeeper que se basa en la coordinación de procesos distribuidos de una forma similar que Hadoop con la diferencia que el procesamiento es en tiempo real entregando como salida un formato en Texto y XML.

6. IBM BigSheets: el formato de entrada pueden ser archivos en formato de valores separados por tabulaciones (TSV) y Hojas de datos, puede almacenar los datos en el sistema de archivos de distribución para su procesamiento y utiliza una interfaz que permite analizar la cantidad de datos y los trabajos de recopilación de larga ejecución mostrando como salida una interfaz gráfica similar a la hoja de Excel.

b) Frameworks:

1. Marimba: La entrada de datos es en tiempo real y el formato de entrada es por medio de la base de datos HBase, el procesamiento es mediante Hadoop y se puede usar para implementar trabajos MapReduce de manera incremental mediante la recalculación de datos y mediante la definición una vista materializada por medio de una consulta declarativa. La salida de datos es mediante la detección de Deltas "Δ" y por recuento incremental de palabras dando como resultado un formato de clave – valor y se puede almacenar en una base de datos HBase.

2. Framework to Handle Data Heterogeneity Contextual (FHDHC - Prototipo): el formato de entrada pueden ser datos obtenidos desde Bases de datos HBase, procesa los datos estructurados centralizados a datos estructurados distribuidos utilizando el entorno basado en Hadoop y mediante un algoritmo se almacena la información en formato Clave – Valor en H-Base.

3. Framework of Integrated Big Data (FIBD): los formatos de entrada pueden ser obtenidas desde tuplas, el procesamiento es mediante la gestión de datos, el análisis y la visualización de datos, realiza operaciones de inserción, eliminación, actualización y

consulta de datos apoyándose de herramientas como Graph, Tensor, HBasse, Hadoop y MapReduce para obtener un modelo de almacén unificado para los 3 tipos de datos con clave – valor, XML, JSON y YAML.

4. BigDAF: el formato de entrada pueden ser datos obtenidos desde Bases de datos HBase, procesa los datos mediante Hadoop y mediante un conjunto de algoritmos para almacenar la información en formato de ficheros distribuidos en HDFS.

5. Dryad y DryadLINQ: los formatos de entrada pueden ser obtenidas desde bases de datos HBase y Cassandra, el procesamiento es mediante Hadoop por medio de clusters para separar y procesar los datos, la salida es un almacén de datos en formato Clave – Valor.

6. Nephele: los formatos de entrada pueden ser obtenidas desde tuplas de bases de datos, el procesamiento es distribuido por medio de Hadoop, Dyrad para obtener un modelo de almacén unificado con clave – valor.

7. Framework (AaaS): los formatos de entrada pueden ser obtenidas desde bases de datos Oracle y DB2, el procesamiento es mediante el uso de algoritmos para llevar un filtrado y etiquetado de datos, así como de un motor semántico para una serie de revisiones para mejorar la calidad de los datos obtenidos con ayuda de herramientas como: Hadoop para procesamiento distribuido, MapReduce, MAhout y R para el procesamiento. Como salida se obtiene un almacén en formato clave valor para su posterior visualización por medio del navegador en forma de tablas y como panel.

8. RUBA: los formatos de entrada pueden ser obtenidas desde bases de datos CEP (Complex Event Processing), el procesamiento es mediante el uso de algoritmos recording, rebuilding y re execution utilizando un motor de procesamiento de eventos complejos para analizar datos en tiempo real. Como salida se obtiene un nuevo almacén CEP.

9. Twister.Net: los formatos de entrada pueden ser obtenidas desde bases de datos, tuplas y hojas de cálculo, el procesamiento es mediante Hadoop y MapReduce para separar la información en un conjunto de datos distribuidos para obtener un almacén de datos con formato Clave – Valor.

c) Metodologías:

1. An Architecture and Methods for Big Data Analysis (AAMBDA):

Presenta un método aplicado a una arquitectura basada en la nube para adquirir, indexar, recopilar, interpretar, procesar, transportar y almacenar datos estructurados, no estructurados y semiestructurados. Para el procesamiento utiliza Hadoop para la indexación y la ubicación de los datos con el fin de presentar al usuario los resultados del análisis de datos en un formato fácil de leer utilizando la minería a través del análisis y la predicción de las evoluciones de datos estructurados, no estructurados y semiestructurados mediante la Integración y unificación de datos.

2. Big Data: Methods, Prospects, Techniques (BDMPT):

Presenta diferentes métodos y técnicas de Big Data para tener una visión general de cómo utilizar el paralelismo y sistemas distribuidos para procesar, gestionar y analizar los distintos tipos de datos, también presentan técnicas que conducen a categorizar dos tipos de tecnologías: procesamiento por lotes y tecnologías de transmisión. Algunos métodos son los siguientes:

• Métodos de minería de Big Data: Paralelismo basado en datos distribuidos y computación en la nube basada en MapReduce.

• Métodos de procesamiento de Big Data: Hashing, Indexación, filtro de floración y computación paralela

3. Multi-model Databases: A New Journey to Handle the Variety of Data (MDANJHVD):

Presenta multimodelos para crear una plataforma de base de datos para administrar la diversidad de los tipos de datos, utilizando modelos relacionales basados en términos matemático de relaciones (subconjuntos) utilizando un DBMS relacional para asegurar el almacenamiento y la recuperación de los datos. Algunos modelos propuestos son: Modelo semiestructurado para documentos XML y JSON, Modelo de clave / valor y el Modelo de grafos dedicada al almacenamiento y la gestión eficiente de los datos.

Anexo B) Descripción para el procesamiento de datos semiestructurados

A continuación, se se muestran las herramientas y frameworks propuestos para procesar los datos semiestructurados de acuerdo al conjunto propuesto para abordar el problema de la variedad (Capitulo 6) y a la Figura 9.

Para el procesamiento de datos semiestructurados se obtuvieron **8 herramientas, 8 frameworks y 4 metodologías** las cuales son:

a) **Herramientas:**

1. Hadoop: el formato de entrada pueden ser archivos en formato de valores separados por tabulaciones (TSV), Snappy, GZip, Deflate, BZip2, CSV, y JSON, el procesamiento distribuido mediante clusters utilizando MapReduce para separar la información en un formato clave – valor y de esta manera almacenar la información en un formato binario dentro de un almacén de datos con formato Clave - Valor.

2. Spark: los formatos de entrada pueden ser obtenidas desde registros de Kafka, Parquet, archivos de secuencia, JSON y CSV, el procesamiento es mediante Hadoop y MapReduce para separar la información en un conjunto de datos distribuidos resilientes(RDD) para distribuir la información, modificarla y obtener alguna salida de ella.

3. Genus: el formato de entrada pueden ser archivos CSV, JSON y XML, el procesamiento es mediante algoritmos de Extracción Transformación y carga para separar la información dando como resultado un almacén de datos NoSQL mediante el lenguaje XML.

4. IBM InfoSphere: el formato de entrada es mediante CSV, XML y Zip, el procesamiento es mediante la integración de datos permitiendo comprender, limpiar, supervisar y transformar datos mediante un lenguaje de consulta declarativa para entregar un formato binario.

5. IBM BigSheets: el formato de entrada pueden ser archivos en formato CSV y JSON, puede almacenar los datos en el sistema de archivos de distribución para su procesamiento y

utiliza una interfaz que permite analizar la cantidad de datos y los trabajos de recopilación de larga ejecución mostrando como salida una interfaz gráfica similar a la hoja de Excel.

6. Project Voldemort: el formato de entrada pueden ser archivos JSON y XML, el procesamiento es distribuido mediante un conjunto de clusters hadoop, los datos se replican automáticamente en varios servidores, el resultado del procesamiento es un sistema de almacenamiento de datos en formato clave – valor.

7. Cloudera Big Data Solutions-(Hadoop): el formato de entrada pueden ser archivos CSV, JSON y datos en formato binario, utiliza Apache hadoop para obtener una salida más valiosa de todos sus datos añadiendo funciones de seguridad, control y gestión mediante un conjunto de herramientas como Sqoop, Flume, Kafka, Hive, apache pig, MR2 y Spark para procesar los datos y obtener un conjunto de archivos en binario para su almacenamiento en formato clave – valor.

8. Samza: La entrada de datos es en tiempo real y el formato de entrada es por medio de Kafka, realiza el procesamiento distribuido utilizando Hadoop y YARN en el proceso, cada tarea contiene un almacén clave-valor usado para almacenar el estado. Como resultado es un almacén en formato de clave – valor.

b) Frameworks:

1. Marimba: La entrada de datos es en tiempo real y el formato de entrada es por medio de CSV, JSON y XML, el procesamiento es mediante Hadoop y se puede usar para implementar trabajos MapReduce de manera incremental mediante la recalculación de datos. La salida de datos es mediante la detección de Deltas "Δ" y por recuento incremental de palabras dando como resultado un formato de clave – valor y se puede almacenar en una base de datos HBase.

2. Framework to Handle Data Heterogeneity Contextual (FHDHC - Prototipo): el formato de entrada es XML, procesa los datos semiestructurados a un formato estructurado utilizando el entorno basado en Hadoop y mediante un algoritmo se almacena la información en formato Clave – Valor en H-Base.

3. Framework of Integrated Big Data (FIBD): el formato de entrada puede ser XML o HTML, el procesamiento es mediante la gestión de datos, el análisis y la visualización de datos, realiza operaciones de inserción, eliminación, actualización y consulta de datos apoyándose de herramientas como Graph, Tensor, HBasse, Hadoop y MapReduce para obtener un modelo de almacén unificado para los 3 tipos de datos con clave – valor, JSON y YAML.

4. BigDAF: el formato de entrada pueden ser XML o JSON, procesa los datos mediante Hadoop y mediante un conjunto de algoritmos para almacenar la información en formato de ficheros distribuidos en HDFS.

5. Dryad y DryadLINQ: el formato de entrada pueden ser XML y HTML, el procesamiento es mediante Hadoop por medio de clúster para separar y procesar los datos, la salida es un almacén de datos en formato Clave – Valor.

6. Nephele: los formatos de entrada pueden ser CSV, JSON y XML, el procesamiento es distribuido por medio de Hadoop, Dyrad para obtener un modelo de almacén unificado con clave – valor.

7. Framework (AaaS): el formato de entrada pueden ser XML, HTML, Formato Clave – Valor y Rich Text Format (RTF), el procesamiento es mediante el uso de algoritmos para llevar un filtrado y etiquetado de datos, así como de un motor semántico para una serie de revisiones para mejorar la calidad de los datos obtenidos con ayuda de herramientas como: Hadoop para procesamiento distribuido, MapReduce, Mahout y R para el procesamiento. Como salida se obtiene un almacén en formato clave valor para su posterior visualización por medio del navegador en forma de tablas y como panel.

8. Framework for Concept-Based Exploration of Semi-Structured (FC-BESSSED): el formato de entrada puede ser JSON, XML, Java y HTML, el procesamiento es mediante el uso de una red conceptual de nube de etiquetas utilizando el procesamiento distribuido por medio de MapReduce. Como salida se obtiene un almacén en formato clave - valor con información basada en etiquetas.

c) Metodologías:

1. An Architecture and Methods for Big Data Analysis (AAMBDA):

Presenta un método aplicado a una arquitectura basada en la nube para adquirir, indexar, recopilar, interpretar, procesar, transportar y almacenar datos estructurados, no estructurados y semiestructurados. Para el procesamiento utiliza Hadoop para la indexación y la ubicación de los datos con el fin de presentar al usuario los resultados del análisis de datos en un formato fácil de leer utilizando la minería a través del análisis y la predicción de las evoluciones de datos estructurados, no estructurados y semiestructurados mediante la Integración y unificación de datos.

2. Big Data: Methods, Prospects, Techniques (BDMPT):

Presenta diferentes métodos y técnicas de Big Data para tener una visión general de cómo utilizar el paralelismo y sistemas distribuidos para procesar, gestionar y analizar los distintos tipos de datos, también presentan técnicas que conducen a categorizar dos tipos de tecnologías: procesamiento por lotes y tecnologías de transmisión. Algunos métodos son los siguientes:

• Métodos de minería de Big Data: Paralelismo basado en datos distribuidos y computación en la nube basada en MapReduce.

• Métodos de procesamiento de Big Data: Hashing, Indexación, filtro de floración y computación paralela

3. Multi-model Databases: A New Journey to Handle the Variety of Data (MDANJHVD):

Presenta multimodelos para crear una plataforma de base de datos para administrar la diversidad de los tipos de datos, utilizando modelos relacionales basados en términos matemático de relaciones (subconjuntos) utilizando un DBMS relacional para asegurar el almacenamiento y la recuperación de los datos. Algunos modelos propuestos son: Modelo semiestructurado para documentos XML y JSON, Modelo de clave / valor y el Modelo de grafos dedicada al almacenamiento y la gestión eficiente de los datos.

4. An Iterative Methodology for Big Data Management, Analysis and Visualization (AIMBDMAV):

Presentan una metodología para abordar proyectos de Big Data de manera sistemática. La metodología es de manera iterativa para la gestión, análisis y visualización de Big Data. Las fuentes de datos que utiliza son Bases de datos NoSQL, para el procesamiento utiliza RDBMS extendido, Hadoop y MapReduce, quedando un almacén de datos en Mongo DB para archivos JSON.

Anexo C) Descripción para el procesamiento de datos no estructurados

A continuación, se se muestran las herramientas y frameworks propuestos para procesar los datos no estructurados de acuerdo al conjunto propuesto para abordar el problema de la variedad (Capitulo 6) y a la Figura 10.

Para el procesamiento de datos no estructurados se obtuvieron **11 herramientas, 13 frameworks y 7 metodologías** las cuales son:

a) **Herramientas:**

1. Hadoop: el formato de entrada pueden ser archivos en formato de texto, el procesamiento distribuido es mediante clúster utilizando MapReduce para separar la información en un formato binario y de esta manera almacenar la información en un formato binario dentro de un almacén de datos con formato Clave - Valor.

2. Spark: el formato de entrada pueden ser archivos en formato de texto, el procesamiento es mediante Hadoop y MapReduce para separar la información en un conjunto de datos distribuidos resilientes(RDD) para distribuir la información, modificarla y obtener alguna salida de ella como lo es CSV.

3. Genus: el formato de entrada pueden ser archivos en formato de texto, imagen y video, el principal objetivo es extraer los datos en estos diferentes tipos de formatos para agruparlos en un almacén de datos coherente, en la mayoría de los casos, este almacén de datos se procesará mediante análisis y algoritmos de extracción de conocimiento. Al considerar los tipos de datos como las imágenes, el almacén de datos resultante será más rico que uno simple construido a partir de datos textuales o numéricos. Otro tipo de datos que influyen en la toma de decisiones son los datos de video. El video podría ser más expresivo que la imagen, deriva su fuerza del aspecto temporal que realiza. Como resultado esperado es un almacén de datos de documentos en una base de datos NoSQL mediante el lenguaje XML.

El principal objetivo es extraer los datos en estos diferentes tipos de formatos para agruparlos en un almacén de datos coherente, en la mayoría de los casos, este almacén de datos se procesará mediante análisis y algoritmos de extracción de conocimiento. Al considerar los

tipos de datos como las imágenes, el almacén de datos resultante será más rico que uno simple construido a partir de datos textuales o numéricos. Otro tipo de datos que influyen en la toma de decisiones son los datos de video. El video podría ser más expresivo que la imagen, deriva su fuerza del aspecto temporal que realiza.

4. IBM InfoSphere: el formato de entrada es texto, el procesamiento es mediante la integración de datos permitiendo comprender, limpiar, supervisar y transformar datos mediante un lenguaje de consulta declarativa para entregar un formato binario.

5. Storm: el formato de entrada es texto, el procesamiento es con ayuda de Apache ZooKeeper que se basa en la coordinación de procesos distribuidos de una forma similar que Hadoop con la diferencia que el procesamiento es en tiempo real entregando como salida un formato en XML.

6. Project Voldemort: el formato de entrada es texto, el procesamiento es distribuido mediante un conjunto de clúster hadoop, los datos se replican automáticamente en varios servidores, el resultado del procesamiento es un sistema de almacenamiento de datos en formato clave – valor.

7. Flink: La entrada de datos es en tiempo real y en lotes, el formato de entrada pueden ser redes sociales, datos de sensores, dispositivos, aplicaciones, registros de mensajes, archivos de sistemas y archivos en formato de texto, el procesamiento distribuido es basado en hadoop en la forma de separar la información en un formato binario mediante HDFS y de esta manera almacenar la información en un almacén de datos con formato Clave - Valor.

8. Mahout: el formato de entrada pueden ser mediante Foursquare como Motor de recomendaciones, Redes sociales para Modelado de intereses de usuario y Yahoo! para Minería de patrones, el procesamiento es basado en la minería de datos basado principalmente en Hadoop extrayendo conocimiento útil a partir de fuentes de datos en bruto apoyándose en la arquitectura Map-Reduce para separar la información en un formato de vectores y un diccionario de términos que luego se podrán utilizar para un agrupamiento en forma clave- valor.

9. Hortonworks: el formato de entrada es texto, el procesamiento de datos es mediante múltiples cargas de trabajo a través de una variedad de métodos de procesamiento y análisis de la información en HDFS y de esta manera almacenar la información en un formato binario en HBase.

10. MapReduce: el formato de entrada es texto, el procesamiento que realiza MapReduce en realidad se refiere a dos procesos separados: El proceso Map toma un conjunto de datos y lo convierte en otro conjunto, donde los elementos individuales son separados en tuplas y el proceso Reduce que obtiene la salida del map como datos de entrada y combina las tupias en un conjunto más pequeño de las misma y de esta manera se obtiene un conjunto de datos en formato binario que posteriormente se pueden almacenar dentro de un almacén de datos con formato Clave - Valor.

11. EpiC: el formato de entrada puede ser texto, registros telefónicos y correos electrónicos, el procesamiento adopta un diseño extensible y soporta dos modelos de procesamiento como MapReduce y el modelo de base de datos de relaciones dividiendo el trabajo analítico en sub tareas y elige los sistemas adecuados para realizar esas sub tareas en función de los tipos de datos para almacenarlos en un formato clave - valor.

b) Frameworks:

1. Marimba: La entrada de datos es en tiempo real y el formato de entrada es por medio de sitios web, datos gráficos, texto y blogs, el procesamiento es mediante Hadoop y se usa para implementar trabajos MapReduce de manera incremental mediante la recalculación de datos definiendo una vista materializada mediante una consulta declarativa teniendo como resultado un formato de clave – valor almacenándolo en una base de datos en HBase.

2. Framework to Handle Data Heterogeneity Contextual (FHDHC - Prototipo): la entrada de datos es por medio de redes sociales y texto, procesa los datos no estructurados a un formato semiestructurado XML utilizando MapReduce para almacenarlos en H-Base.

3. Framework of Integrated Big Data (FIBD): el formato de entrada puede ser datos gráficos, texto, audio e imagen, el procesamiento es mediante la gestión de datos, el análisis

y la visualización de datos, realiza operaciones de inserción, eliminación, actualización y consulta de datos apoyándose de herramientas como Graph, Tensor, HBasse, Hadoop y MapReduce para obtener un modelo de almacén unificado para los 3 tipos de datos con clave – valor, JSON y YAML.

4. BigDAF: el formato de entrada pueden ser texto, sensores, dispositivos y blogs, procesa los datos mediante Hadoop y mediante un conjunto de algoritmos para almacenar la información en formato de ficheros distribuidos en HDFS.

5. Dryad y DryadLINQ: el formato de entrada es texto, el procesamiento es mediante Hadoop por medio de clúster para separar y procesar los datos, la salida es un almacén de datos en formato Clave – Valor.

6. Nephele: la entrada de datos es por medio del procesamiento en la nube, el procesamiento lo realiza de manera distribuida por medio de Hadoop y mediante Dyrad se puede obtener un modelo de almacén unificado con clave – valor.

7. Framework (AaaS): el formato de entrada pueden ser texto y web semántica, el procesamiento es mediante el uso de algoritmos para llevar un filtrado y etiquetado de datos, así como de un motor semántico para una serie de revisiones para mejorar la calidad de los datos obtenidos con ayuda de herramientas como: Hadoop para procesamiento distribuido, MapReduce, Mahout y R para el procesamiento. Como salida se obtiene un almacén en formato clave valor para su posterior visualización por medio del navegador en forma de tablas y como panel y de esta manera extraer mayor información de los datos no estructurados.

8. RUBA (Framework propuesto): el formato de entrada puede ser imagen, video, Circuito cerrado de televisión (CCTV) y sistemas de monitorización, utiliza un motor de procesamiento de eventos complejos para analizar datos en tiempo reale mediante el uso de algoritmos recording, rebuilding y re execution convirtiendo los datos no estructurados en datos estructurados. Como salida se obtiene un nuevo almacén CEP.

9. Framework for Unstructured Data Analysis (FUDA – Frameworks en desarrollo): el formato de entrada puede ser texto y redes sociales, el procesamiento de los datos es mediante el clúster Hadoop usando Hbase para almacenar los datos después de un análisis, se utiliza Java para interactuar con el servidor y posteriormente se obtendrá como formato resultante un XML.

10. Framework ETL: el formato de entrada es mediante la web semántica, para el procesamiento utiliza tecnologías semánticas para conectar, vincular y cargar datos en un almacén de datos, a partir de los datos semánticos se realiza otro procesamiento utilizando RDF como modelo de datos gráficos junto con SPARQL como lenguaje de consulta semántica y de esta manera obtener una extracción de datos en formatos de archivo plano, para la transformación se utilizan algoritmos de estructuración y limpieza de datos, por último la carga de datos implica la propagación de los datos en un almacén de datos en formato clave – valor en HBase.

11. jMetalSP: el formato de entrada es mediante la web semántica, optimiza los datos no estructurados en tiempo real mediante librerías de algoritmos de optimización adaptados para los datos heterogéneos, el procesamiento es por medio de Apache Spark separando la información en un conjunto de datos distribuidos resilientes(RDD) para analizar la información, modificarla y obtener alguna salida de ella como lo es un formato en clave – valor para almacenarlo en HBase.

12. Piccolo: el formato de entrada puede ser sitios web, datos gráficos y texto, el procesamiento es promedio de un modelo de programación centrado en datos para escribir aplicaciones paralelas consistiendo en funciones de control, que se ejecutan en una sola máquina, y funciones del núcleo, que se ejecutan simultáneamente en muchas máquinas para obtener un almacén en formato clave – valor.

13. Framework for Extracting Reliable Information from Unstructured Uncertain Big Data (FERIUUBD - prototipo): el formato de entrada puede ser sitios web, texto y audio, para el procesamiento se manejan dos enfoques para manejar los tipos de datos no estructurados e inciertos. La forma principal es utilizar un enfoque de combinación de información, que se une a varias fuentes menos confiables para hacer más útil los datos, y el

segundo camino es a través de la lógica difusa, enfoques de conjuntos suaves y técnicas de optimización robustas. Para lograrlo se realizan cálculos creados en la etapa Hadoop para usar la computación distribuida y de esta manera extraer información para almacenarla en un formato clave – valor.

c) Metodologías:

1. An Architecture and Methods for Big Data Analysis (AAMBDA):

Presenta un método aplicado a una arquitectura basada en la nube para adquirir, indexar, recopilar, interpretar, procesar, transportar y almacenar datos estructurados, no estructurados y semiestructurados. Para el procesamiento utiliza Hadoop para la indexación y la ubicación de los datos con el fin de presentar al usuario los resultados del análisis de datos en un formato fácil de leer utilizando la minería a través del análisis y la predicción de las evoluciones de datos estructurados, no estructurados y semiestructurados mediante la Integración y unificación de datos.

2. Big Data: Methods, Prospects, Techniques (BDMPT):

Presenta diferentes métodos y técnicas de Big Data para tener una visión general de cómo utilizar el paralelismo y sistemas distribuidos para procesar, gestionar y analizar los distintos tipos de datos, también presentan técnicas que conducen a categorizar dos tipos de tecnologías: procesamiento por lotes y tecnologías de transmisión. Algunos métodos son los siguientes:

• Métodos de minería de Big Data: Paralelismo basado en datos distribuidos y computación en la nube basada en MapReduce.

• Métodos de procesamiento de Big Data: Hashing, Indexación, filtro de floración y computación paralela

3. Multi-model Databases: A New Journey to Handle the Variety of Data (MDANJHVD):

Presenta multimodelos para crear una plataforma de base de datos para administrar la diversidad de los tipos de datos, utilizando modelos relacionales basados en términos matemático de relaciones (subconjuntos) utilizando un DBMS relacional para asegurar el almacenamiento y la recuperación de los datos. Algunos modelos propuestos son: Modelo semiestructurado para documentos XML y JSON, Modelo de clave / valor y el Modelo de grafos dedicada al almacenamiento y la gestión eficiente de los datos.

4. An Iterative Methodology for Big Data Management, Analysis and Visualization (AIMBDMAV):

Presentan una metodología para abordar proyectos de Big Data de manera sistemática. La metodología es de manera iterativa para la gestión, análisis y visualización de Big Data. Las fuentes de datos que utiliza son Bases de datos NoSQL, para el procesamiento utiliza RDBMS extendido, Hadoop y MapReduce, quedando un almacén de datos en Mongo DB para archivos JSON.

5. A Storage Model for Handling Big Data Variety (ASMHBDV):

El modelo propuesto toma una decisión de acuerdo a la asignación de recursos de almacenamiento (Bases de datos) en función del tipo de datos. Para ello utilizan distintos métodos como:

I. La recopilación de información relacionada con el almacenamiento de datos controlados por bases de datos disponibles como Cassandra, SQL, MongoDB, Neo4j, HDFS.

II. La organización de la información de almacenamiento en formato XML.
Como resultado se propone un modelo de almacenamiento basado en XML que resolvería el problema de almacenar cualquier tipo de datos.

6. Beyond the hype: Big Data concepts, methods, and analytics (BHBDCMA):

Describe métodos analíticos para el análisis de datos no estructurados en formatos de texto, audio y video. Los métodos para cada tipo de dato son los siguientes:

1) Análisis de texto

La analítica de texto implica análisis estadístico, lingüística computacional y aprendizaje automático. Algunas técnicas de extracción de información (IE) son el reconocimiento de entidades (ER) y la extracción de relaciones (RE). ER encuentra nombres en el texto y los clasifica en categorías predefinidas. RE encuentra y extrae relaciones semánticas entre entidades en el texto.

2) Análisis de audio

El análisis de audio analiza y extrae información de datos de audio no estructurados. La analítica del habla sigue dos enfoques tecnológicos comunes: el enfoque basado en la transcripción (conocido como reconocimiento de voz continuo de gran vocabulario, LVCSR) y el enfoque basado en la fonética.

3) Analítica de video

La analítica de video involucra una variedad de técnicas para monitorear, analizar y extraer información significativa de las transmisiones de video, el procesamiento de un video es por medio de metadatos, banda sonora, transcripciones y el contenido visual.

4) Analítica predictiva

El análisis predictivo comprende una variedad de técnicas que predicen resultados futuros basados en datos históricos y actuales. Las técnicas de análisis predictivo se basan principalmente en métodos estadísticos.

7. Big Data preprocessing: methods and prospects (BDPMP - IMAGS):

Presentan métodos de pre procesamiento de datos para la minería de datos en Big Data. El procesamiento se realiza mediante la transformación, integración, limpieza y normalización de los datos.

Se utiliza principalmente para minería de texto que intentan estructurar el texto de entrada, produciendo patrones estructurados de información. Utiliza una variedad de herramientas para el análisis de la información como:

- TF-IDF: esta herramienta cuantifica la relevancia de cada término para un documento, dado un conjunto completo de documento midiendo la cantidad de veces que aparece un término en un documento, así como su frecuencia.
- Word2Vec: toma como entrada un corpus de texto y produce como salida los vectores de palabras. Primero construye un vocabulario del texto y luego aprende la representación vectorial de las palabras.

- CountVectorizer: transforma un corpus en un conjunto de vectores de tokens. Extrae el vocabulario usando un estimador y cuenta el número de ocurrencias para cada término.

- Tokenizer: divide algunos textos en términos individuales utilizando expresiones simples o regulares.

- StopWordsRemover: elimina palabras irrelevantes del texto de entrada. La lista de palabras de detención se especifica como parámetro.

- n-gram: genera secuencias de términos n-gramos, donde cada uno está formado por una cadena delimitada por espacios de n palabras consecutivas.

Todo el análisis del texto puede ser almacenado en un formato de clave – valor para su extracción de conocimiento.