ESTUDIO DE LAS TÉCNICAS DE ANÁLISIS BIG DATA PARA DATOS HIDROCLIMATOLÓGICOS DE UNA CUENCA HÍDRICA

KELLY TATIANA LÓPEZ TANGARIFE
GLORIA LORENA SANDOVAL GALLEGO

UNIVERSIDAD TECNOLÓGICA DE PEREIRA.
FACULTAD DE INGENIERÍAS.
INGENIERÍA DE SISTEMAS Y COMPUTACIÓN.
PEREIRA.
2017

ESTUDIO DE LAS TÉCNICAS DE ANÁLISIS BIG DATA PARA DATOS HIDROCLIMATOLÓGICOS DE UNA CUENCA HÍDRICA

Trabajo de Grado para Optar al Título de
Ingeniero de Sistemas y Computación.

KELLY TATIANA LÓPEZ TANGARIFE
GLORIA LORENA SANDOVAL GALLEGO

Director.
Ing. JORGE IVÁN RÍOS PATIÑO

UNIVERSIDAD TECNOLÓGICA DE PEREIRA.
FACULTAD DE INGENIERÍAS.
INGENIERÍA DE SISTEMAS Y COMPUTACIÓN.
PEREIRA.
2017

A nuestros padres y seres amados.

Agradecimientos

No alcanzan las palabras para expresar la gratitud que sentimos primero hacia Dios por darnos la vida y las fuerzas de luchar por alcanzar nuestros sueños y segundo a nuestros padres que han sido un apoyo incondicional en esta carrera que emprendimos. Gracias a ese amor entrañable con el que nos han formado, hoy somos lo que somos.

A nuestros hermanos y amigos, es un privilegio tenerlos con nosotros pues siempre nos alientan con sus consejos y nos llena la vida de sonrisas y alegrías.

A todas aquellas personas que en diferentes circunstacias nos tendieron su mano cuando más lo necesitamos, esperando únicamente algo a cambio: vernos triunfar y lograr nuestros objetivos, a ellos mil gracias.

Al ingeniero Jorge Iván Ríos por guiarnos en la realización de este proyecto y por su entrega y compromiso a la hora de compartir el conocimiento adquirido.

A todos los docentes que día a día nos impregnaron de su conocimiento y nos compartieron sus vivencias con el fin de formar profesionales integrales, todo nuestro aprecio.

Índice general

Índice de figuras

Índice de cuadros

1

Introducción

El avance de las tecnologías de la comunicación y la computación, han sido el elemento clave en el desarrollo de la tecnología de internet, cada vez más personas tienen acceso a este debido a su alto nivel de escalamiento, lo que ha provocado que se convierta en la plataforma de todos los tipos de interacción humana [1]. A través de esta interacción son generados todos los datos suficientes para que una organización analice y posteriormente diseñe mejores estrategias de marketing, optimice sus procesos, tenga ventajas competitivas, fortalezca la toma de decisiones, entre otros beneficios importantes.

El auge que se ha venido presentando en cuanto a generación de datos, que no están precisamente en un formato o estructura única, sino que provienen de diversas fuentes, dando origen a datos estructurados, no estructurados y semiestructurados (texto, fotos, audio, por ejemplo), es lo que ha causado el uso de la arquitectura Big Data. Se estima que sólo el 5 % de los datos digitales, son estructurados [2], lo que da a entender que el 95 % de datos restante se compone tanto de datos no estructurados, como semiestructurados.

El valor que dichos datos tienen para una organización es imprescindible, pero no saber obtenerlo se convierte en una desventaja competitiva puesto que un pequeño grupo de pioneros ha empezado a comprender y explorar cómo procesar y analizar de nuevas formas la gran variedad de información, logrando resultados empresariales muy importantes [3], así como sociales y medioambientales.

En nuestro caso, donde el foco de análisis serán las variables tomadas de una cuenca hídrica, es preciso resaltar que cada 5 minutos (según Aguas & Aguas) se están midiendo a través de las estaciones de telemetría, variables hidroclimatológicas como velocidad y dirección del viento, pluviosidad, presión

atmosférica, presión barométrica, humedad relativa, radiación solar, batimetría de la cuenca, volumen y velocidad del cauce, entre otras, que luego de ser analizadas en su conjunto, permiten conocer el estado de la cuenca hídrica y tener cierto grado de confianza para tomar decisiones como prevenir inundaciones, realizar la ordenación del territorio alrededor de los ríos, exigir criterios de diseño de obras e infraestructuras que puedan funcionar satisfactoriamente en situaciones de emergencia y otras decisiones al respecto.

A través del presente trabajo de investigación formativa se hará alusión a las técnicas de análisis Big Data para sentar las bases en lo que tiene que ver con el análisis de datos hidroclimatológicos de una cuenca hídrica para lo cual se hablará de los aspectos fundamentales del término como son su precedente histórico, las definiciones que se conocen al respecto y las características que lo definen. Además se hará un recorrido por algunas técnicas de análisis, herramientas y tecnologías usadas en la actualidad por las organizaciones para sacar provecho de los datos que almacenan día a día. A su vez, se hará una contextualización sobre lo que es la hidroclimatología, las variables principales a la hora de analizar este tipo de datos y como Big Data juega un papel fundamental en este proceso.

2

Metodología

Se realizará la investigación basada en los datos bibliográficos como un proceso secuencial, de recolección, clasificación y evaluación de la información tanto física como virtual, que servirá como base fundamental para la fuente de teoría que forma parte de investigación en el desarrollo de la monografía.

La investigación se realizará mediante los siguientes pasos:

- Basados en los objetivos establecidos en la investigación: se hará un proceso de recolección de información, de fuentes bibliográficas y artículos tanto físicos como virtuales que ayudarán al pertinente desarrollo de la investigación acerca de las técnicas de analisis Big Data e hidroclimatología.
- Las fuentes bibliográficas encontradas se clasificarán dependiendo del orden en que se va a desarrollar el proceso de investigación, además del tiempo que llevan de ser publicados.
- Se define la forma del registro de las fuentes de consulta, basándose en las referencias bibliográficas de las normas ISO.

Se tendrá en cuenta sólo información que provenga de fuentes confiables y que su tiempo de publicación sea menor a cinco años, teniendo como objetivo principal sustentar todos los datos que se mencionan en la monografía; además se hará una investigación descriptiva que comprende la definición, análisis e interpretación del tema que se está desarrollando.

3

Generalidades de Big Data

3.1. Antecedentes históricos del término Big Data

El termino Big Data tiene un gran precedente histórico que surge tras la necesidad de darle una denominación a la velocidad de crecimiento que empezó a tener la información, lo que fue conocido en 1941 como "La explosión de la información" [4] de acuerdo al Diccionario Inglés de Oxford.

Varios autores [5] (como por ejemplo Gil Press) ofrecen una mirada a través de la historia sobre la explosión de los datos, de una manera clara y precisa que nos permite comprender cómo se sentaron los precedentes para lo que hoy conocemos como Big Data. Los hitos principales que serán citados a continuación abarcan desde finales siglo XVIII hasta la época actual:

- **1875:** Se trata de darle definición al término Big Data que será de gran relevancia para la utilización del mismo a lo largo de todos los años.
- **1880:** Se empieza a dar la primera sobrecarga de información en el censo de los estados unidos al tratar de procesar la información colectada, dicho proceso tardo más de 8 años en completarse con los métodos de tabulación existentes de la época.
- **1932:** Se dio la sobrecarga de información con el crecimiento desmesurado de la población de los estados unidos.

- **1944:** Se estima que las bibliotecas universitarias estadounidenses duplicaban su tamaño cada diez y seis años. Por lo cual se necesitaría más espacio para almacenar esa cantidad de libros.
- **1948:** Claude Shannon publicó la Teoría Matemática de la Comunicación, en la que se estableció un marco de trabajo para determinar los requisitos de datos mínimos para transmitir la información a través de canales afectados por ruido (imperfectos).
- **1956:** Se creó el concepto de memoria virtual realizada por el físico alemán Fritz Rudolf, con la idea de almacenamiento finito e infinito. El almacenamiento de datos administrado por el hardware y el software integrado, para esconder los detalles del usuario, permitió el procesamiento de datos sin tener las limitaciones de memoria de hardware.
- **1961:** El científico Derek Price, crea la revolución científica que es la encargada tanto de la comunicación rápida de ideas como de la información científica. Concluye que el número de revistas nuevas ha crecido exponencialmente en lugar de hacerlo de forma lineal, duplicándose cada quince años y aumentando en un factor de diez cada medio siglo.
- **1962:** Se empezaron a hacer los primeros desarrollos para el reconocimiento de voz y se crea la primera máquina que reconocía 16 palabras en inglés y era capaz de procesarlos correctamente.
- **1966:** Empiezan a surgir los sistemas centralizados de cómputo.
- **1967:** BA Marron y PAD de Maine publican "Compresión automática de datos" en las Comunicaciones de la ACM, donde se habla sobre la "explosión de la información". En el documento se describen formas más eficientes para la compresión de información con el objetivo de almacenarla más eficientemente.
- **1970:** Edgar F. Codd, un matemático formado en Oxford, escribió un artículo donde se explica cómo acceder de forma más rápida a una base de datos que tiene una gran cantidad de información, sin saber cómo estaba estructurada [6].
- **1975:** El Ministerio de Correos y Telecomunicaciones de Japón realizó el Censo del Flujo de Información, donde se buscaba obtener evidencia empírica de la cantidad y los medios de información en circulación en la sociedad Japonesa, para lo cual tomaron como unidad de medida "el número de palabras" [7].

- **1980:** IA Tjomsland da una charla titulada "¿A dónde vamos?", donde afirmó que «aquellos que trabajan en dispositivos de almacenamiento descubrieron hace mucho tiempo que la primera ley de Parkinson puede parafrasearse para describir nuestro sector: "los datos se expanden para llenar el espacio disponible" ». Esto debido a que las personas no tienen la forma de identificar datos obsoletos, pues se considera más grave descartar datos que pueden ser potencialmente usados, que retenerlos.

- **1983:** Los avances tecnológicos permitieron a todos los sectores beneficiarse de nuevas maneras de organizar, almacenar y generar datos. Las empresas estaban empezando a usar los datos para tomar mejores decisiones de negocio.

- **1986:** Hal B. Becker publica "¿Pueden los usuarios realmente absorber datos a las velocidades de hoy? A las de Mañana?" en las comunicaciones de datos. Allí mencionaba que «la densidad de recodificación lograda por Gutenberg fue aproximadamente de 500 símbolos (caracteres) por pulgada cúbica; 500 veces la densidad de las tablillas de barro [Sumeria del año 4000 antes de cristo].

- **1990:** Peter J. Denning publica "Almacenando todos los bits", donde se afirma que guardar todos los datos, nos pone en una situación imposible debido a que el volumen y la velocidad abruman las redes, los dispositivos de almacenamiento, la capacidad de comprensión de los humanos, entre otros, lo que conlleva a preguntarse si es posible construir una máquina que pueda soportar toda esta gran cantidad de información.

- **1992:** Crystal Reports creó el primer informe de base de datos sencillo con Windows. Estos informes permitían a las empresas crear un informe sencillo a partir de diversos orígenes de datos con escasa programación de código.

- **1997:** Se usó el término Big Data en Julio de 1997 en el artículo "Application-Controlled Demand Paging for Out-of-Core Visualization" de investigadores de la NASA. Michael Cox y David Ellsworth, afirmaban que el aumento de los datos se estaba convirtiendo en un reto interesante para los sistemas informáticos, dado que agotaban las capacidades de la memoria principal, el disco duro, y aún de los discos remotos [8]. Ellos le atribuyeron a este suceso el nombre "el problema de Big Data", mencionando que cuando los conjuntos de datos no caben en la memoria principal, o cuando ellos no encajan en el disco local, la solución más común es adquirir más recursos. Si bien este es uno de los usos del término más aceptado por ejemplo por Steve Lohr [9], hay otras visiones como la de Bernard Marr, cuya descripción se

enfoca más en las fundaciones históricas de Big Data, afirmando que ayuda a los humanos a capturar, almacenar, analizar y recuperar tanto datos como información. Marr cree que quien por primera vez utilizó el término Big Data fue Erik Larson, debido a que presentó un artículo para la revista "Happer" en cuyo título había dos frases en las que estaban las palabras Big Data. Aun así Lohr defiende su posición, puesto que para él la definición que le dieron al término Michael Cox y David Ellsworth, tenía una precisión razonable, dada la connotación que tiene el término en la actualidad.

- **1999:** El término "Internet de las cosas" o IoT, por sus siglas en inglés, fue acuñado por el emprendedor británico Kevin Ashton, cofundador del Auto-ID Center del MIT, durante una presentación que enlazaba la idea de identificación por radiofrecuencia (RFID) en la cadena de suministro con el mundo de Internet [10].

- **2000:** Debido al auge del crecimiento de los datos, varios investigadores, Peter Lyman y Hal R. Varian de la Universidad Berkeley de California, sintieron la necesidad de cuantificar la información total nueva y original creada anualmente en el mundo [5], en términos de almacenamiento informático. Este estudio fue titulado "How Much Information?" que se publicó en el año 2000 y que tuvo una actualización en el 2003.

- **2001:** Doug Laney, analista de Gartner, publicó un artículo titulado "3D Data Management: Controlling Data Volume, Velocity, and Variety". Al día de hoy, las tres V siguen siendo las dimensiones comúnmente aceptadas de Big Data.

- **2005:** Tim O'Reilly publicó "What is Web 2.0?", donde afirma que «los datos son el próximo Intel Inside©».

- **2007:** En marzo de 2007 investigadores de la corporación de datos Internacional (International Data Corporation), publicaron un artículo titulado "The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010", donde estimaron y proyectaron la cantidad de datos digitales que serían creados y reproducidos cada año.

- **2008:** Bret Swanson y George Gilder publican "Estimating the Exaflood (PDF)", en el cual proyectan que el tráfico IP de Estados Unidos podría alcanzar un zettabyte para 2015 y que la Internet de los Estados Unidos de 2015 sería al menos 50 veces mayor que en 2006 [11].

- **2009:** La inteligencia de negocios llega a ser una prioridad para los gerentes de TI [12].

- **2010:** The Economist publicó el informe titulado "Data, Data Everywhere". En él, su autor Kenneth Cukier escribe: el mundo contiene una cantidad de información inimaginable, cuyo ritmo de crecimiento es extraordinariamente grande.
- **2011:** En un artículo titulado "The World's Technological Capacity to Store, Communicate, and Compute Information" de Science Magazine, se calculó que la capacidad mundial de almacenamiento de información creció a una tasa del 25 % anual desde 1987 hasta 2007.
- **2012:** El artículo "Critical Questions for Big Data", publicado en Information, Communications, and Society Journal, define el término Big Data como «un fenómeno cultural, tecnológico e intelectual que aparece por la interconexión de los siguientes elementos: tecnología, análisis y mitología».
- **2013:** Avances tecnológicos en alza en 2013: (1) Capacidad de realizar consultas federadas, que ofrecen al usuario la posibilidad de adoptar una consulta y aportar soluciones basadas en información procedente de numerosas fuentes diferentes, y (2) capacidad de generar informes a partir de bases de datos almacenadas en memoria, que ofrecen un rendimiento más rápido y más predecible [13].
- **2014:** El IoT (Internet de las Cosas) se ha convertido en una fuerza poderosa para la transformación de negocios y su enorme impacto afectará en los próximos años a todos los sectores y todas las áreas de la sociedad. Existen enormes redes de objetos físicos dedicados (cosas) que incorporan tecnología para detectar o interactuar con su estado interno o medio externo. Según Gartner, había 3,7 billones de "cosas" conectadas en uso en 2014 y esa cifra se elevará hasta los 4,9 billones en 2015 [12] incluyendo sistemas de iluminado LED inteligentes, de monitorización de salud, cerraduras inteligentes y numerosas redes de sensores para detección de movimiento, estudio de contaminación atmosférica, etc.
- **2015:** Una ciudad inteligente (smart city) hace uso del análisis de información contextual en tiempo real para mejorar la calidad y el rendimiento de los servicios urbanos, reducir costes, optimizar recursos e interactuar de forma activa con los ciudadanos.

- **2016:** El IoT, la computación en la nube y Big Data convergen para ofrecer a las empresas una gran oportunidad de analizar sus datos, además como señala Bernard Marr en [14] una estrategia Big Data más que ser atractiva, debe ser práctica con el fin último de encontrar esos patrones y tendencias que se desean descubrir a través de los datos. En este año, las organizaciones movieron cada vez más los proyectos de análisis de datos a la producción, buscando la capacidad de interrogar mejor los datos internos y externos para comprender mejor a sus clientes y aumentar la eficiencia.

3.2. Definiciones de Big Data

Ha ocurrido un cambio en la mentalidad sobre cómo los datos deben ser usados: ya no son considerados como estáticos o "viejos" cuya utilidad llega a su fin en el momento en que se alcanza el propósito para el cual fueron colectados. Por el contrario son estimados como un material crudo de los negocios, un recurso económico vital, usado para crear una nueva forma de valor económico. Usar esos datos inteligentemente, permite incluso tener una fuente de innovación y nuevos servicios [2].

Además, no solo es el mundo inundado de más información que antes, sino que esa información tiene una tasa de crecimiento ingente. Según Viktor Mayer-Schönberger y Kenneth Cukier las ciencias como la astronomía y la genética que experimentaron por primera vez la explosión de datos en el año 2000, acuñaron el término "Big Data", afirmando que el concepto está migrando a todas las áreas que comprende la actividad humana. Sin embargo, como se podrá observar a través de este trabajo de investigación, no se posee una definición precisa ni un origen común para el término, es por eso que se hará alusión a algunas definiciones.

La idea inicial alrededor de Big Data, consistió en que el volumen de información ha crecido tanto que la cantidad analizada no podía ser contenida en la memoria que los computadores usan para el procesamiento, ocasionando la modernización de las herramientas que los ingenieros utilizaban para analizarlos en su totalidad. Es así como se comienza a hablar de nuevas tecnologías de procesamiento como MapReduce de Google y su software equivalente Hadoop que fue lanzado en 2008 por Yahoo como un proyecto de código abierto. El desarrollo de las nuevas herramientas permite la administración de muchos más datos que antes, requiriendo que éstos, sean almacenados de formas diferentes a las de antaño, donde se ubicaban en filas ordenadas o tablas de bases de datos clásicas.

Para Viktor Mayer-Schönberger y Kenneth Cukier Big Data se refiere a las cosas que se pueden hacer a gran escala y que no pueden ser realizadas a una inferior, con el objetivo de extraer nuevos ïnsights"(en español, traduce percepciones, pero no es preciso), o crear nuevas formas de valor, de manera que ocurra un cambio en los mercados, las organizaciones, las relaciones entre los ciudadanos y los gobiernos, y más. Ellos sostienen que es sólo el comienzo, dado que Big Data desafía la forma en la cual vivimos e interactuamos con el mundo; sin embargo, el término tiene otras definiciones y connotaciones más amplias, de acuerdo a otros autores.

En el artículo de IBM titulado ¿Qué es Big Data? [15], al definir el término, expresan: "en términos generales podríamos referirnos como a la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos".

Esa información, que está en un gran volumen, existe además en una gran variedad de datos que puede ser representada de múltiples maneras, donde dichos datos adquieren la forma de mensajes, actualizaciones e imágenes publicadas en las redes sociales; lecturas de sensores; señales GPS de los teléfonos celulares, etcétera [16], donde se mide por ejemplo la temperatura, la humedad, el movimiento, el posicionamiento, entre otras muchas variables que se pueden llegar a cuantificar, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea suficientemente rápida con el fin de obtener la información necesaria en el momento preciso.

De manera semejante, Amazon [17], describe el término en torno a los retos de administración de datos que, debido al incremento en el volumen, la velocidad y la variedad de los mismos, no se puede resolver con las bases de datos tradicionales. En este punto, tanto IBM como Amazon, coinciden en el uso de las 3V's de Big data, como las características principales que se deben tener en cuenta al hablar del término. Otra definición encontrada es la del Diccionario Inglés de Oxford [4] (OED por sus siglas en inglés): "los datos de un tamaño colosal que, por lo general en el grado de su manipulación y gestión presentan desafíos logísticos significativos".

Por parte de Wikipedia, Big Data es "un término para los conjuntos de datos que son tan grandes o complejos que las aplicaciones tradicionales de tratamiento de datos son inadecuadas para darles tratamiento. Los desafíos incluyen análisis, captura, limpieza de datos, búsqueda, intercambio, almacenamiento, transferencia, visualización, consulta, actualización y privacidad de la información" [18].

El estudio publicado en Junio de 2011 por McKinsey Global Institute (MGI) [19] destacó el desafío que trae consigo la definición de la expresión en mención, definiendo Big Data como "conjuntos de datos cuyo tamaño va más allá de la capacidad de las herramientas típicas de software de base de datos para capturar, almacenar, gestionar y analizar", donde se hace relación no tanto al tamaño de los datos explícitamente medidos en petabytes o exabytes, sino al tamaño de esos datos ligado al constante avance de la tecnología. En este caso lo que se pone de relieve es el contexto de Big Data.

En cuanto a Microsoft en [20] "Big Data es el término cada vez más usado para describir el proceso de aplicar un significativo poder de cómputo - el último en aprendizaje de máquina e inteligencia artificial – para conjuntos de información que son con frecuencia altamente complejos y extremadamente grandes".

Para John Akred [21], director de tecnología de Silicon Valley Data Science, "Big Data se refiere a la combinación de un enfoque para informar la toma de decisiones con percepción/entendimiento analítico derivado de los datos, y un conjunto de tecnologías facultativas que permiten que esas percepciones sean derivadas económicamente a partir de, ocasionalmente, diversas fuentes de datos muy grandes. Los avances en las tecnologías sensoriales, la digitalización del comercio y las comunicaciones, y la aparición y crecimiento de los medios sociales son algunas de las tendencias que han creado la oportunidad de utilizar a gran escala, datos de grano fino para entender los sistemas, el comportamiento y el comercio; mientras la innovación en la tecnología hace que sea económicamente viable utilizar esa información para tomar decisiones informadas y mejorar los resultados".

Según Daniel Gillick, investigador científico de Google, "Históricamente, la mayoría de las decisiones – políticas, militares, empresariales y personales – han sido tomadas por cerebros que tienen lógica impredecible y operan sobre evidencia experimental subjetiva. "Big data" representa un cambio cultural en el cual más y más decisiones son tomadas por algoritmos con lógica transparente,

operando sobre evidencia documentada inmutable. Yo pienso que "big" se refiere más a la naturaleza persuasiva de este cambio que a alguna cantidad de datos en particular" [21].

Son muchos los conceptos que se pueden encontrar de Big Data, lo cual ratifica que es un término con un "objetivo en movimiento" como se cataloga en la actualidad; no obstante, aunque varios autores no relacionan la palabra "big" con el tamaño o cantidad de datos, hay una gran parte de ellos que si lo hacen y como se ha podido apreciar, palabras como "grande" y "tradicional", son ambiguas, porque lo "grande" de hoy, será lo "pequeño" del mañana, y algo similar ocurre con lo "tradicional", puesto que las herramientas "nuevas o actuales" tarde que temprano, se convertirán en tradicionales, debido a la constante evolución de la tecnología. Es por eso, que para dar una definición un poco más clara, se va a realizar un enfoque sobre las V's de Big Data.

Desde 1997 muchos atributos se han añadido al término; en medio de ellos, tres se caracterizan por ser los más populares como se pudo ver en las definiciones previas, los cuales han sido los más aceptados y adoptados por los autores. Estas 3V's las propuso Doug Laney en 2001 en un artículo publicado por Meta Group [22], en el cual afirma que el comercio electrónico "E-Commerce" ha explotado los desafíos de la administración de los datos a lo largo de tres dimensiones a saber, el volumen, la velocidad y la variedad. Cada transacción comercial que se realiza a través de canales electrónicos, incrementa la profundidad/amplitud de los datos disponibles; aumenta la velocidad del punto de interacción (intercambio de información entre personas y ordenadores) y, en consecuencia el paso de datos usados para soportar las interacciones y los generados por dichas interacciones. A su vez el autor establece la mayor barrera para la administración efectiva de los datos, debido a la variedad de formatos de datos incompatibles, las estructuras de datos no alineadas y las semánticas de datos inconsistentes, puesto que según Laney, no existirá un mayor obstáculo que dicha variedad.

Posteriormente, IBM [23] agregó un cuarto atributo llamado "veracidad" a las 3Vs establecidas por Laney, donde el volumen representa la escala de datos, la velocidad denota el análisis del flujo de datos, la variedad indica las diferentes formas de los datos y la veracidad implica la incertidumbre de los mismos. Ellos incluyeron ésta última característica dado que se relaciona con la confiabilidad o la falta de esta en los datos y debido a que la variedad de estos hace que la calidad y la precisión sea menos controlable. Subsiguiente a la adición del cuarto atributo, IBM llega a mencionar un quinto: el valor [24]. Esta característica es muy importante en las organizaciones empresariales ya que da la capacidad de alcanzar un mayor valor a partir del análisis de datos a volúmenes, velocidades, variedades o veracidades superiores.

Se ha mencionado el caso de IBM por dar un referente, pero en realidad muchas organizaciones y autores han hecho mención de los tres y más atributos de Big Data. Es el caso por ejemplo, de Bernard Marr en [25] que actualmente trabaja en el Instituto de Desarrollo Avanzado (Avanced Performance Institute) y define las 5V's mencionadas; Ben Larson [26] quien habla sobre las 6V's donde la sexta es "valencia"que se refiere a la habilidad de un dato para combinarse con otros; Mark van Rijmenam fundador de Datafloq, quien a partir de las 3V's dadas por Laney incorpora otras cuatro: veracidad, variabilidad, visualización y valor, es decir, define las 7V's de Big Data; Bill Vorhies [27] quien agrega dos nuevos atributos a saber, viscosidad y viralidad, llegando así a las 8V's; Suhail Sami Owais y Nada Sael Hussein quienes en [28] nombran una novena: la "volatilidad"; Kirk Borne quien en [29] cita 10V's y finalmente Richard Self de la Universidad de Derby [30] quien llega incluso a proponer las 12 V's de Big Data.

Para concluir, una definición bastante completa que se pudo apreciar a través de la investigación fue la que realizaron los autores del libro "Big Data: Principles and Paradigms" [9, Pag. 14], que posterior a un análisis hecho a partir de la definición de las 3^2 V's de Big Data desde tres perspectivas diferentes, terminaron por dar una interpretación suficientemente comprensiva del término. Para ello, usaron el diagrama de Venn como se muestra a continuación:

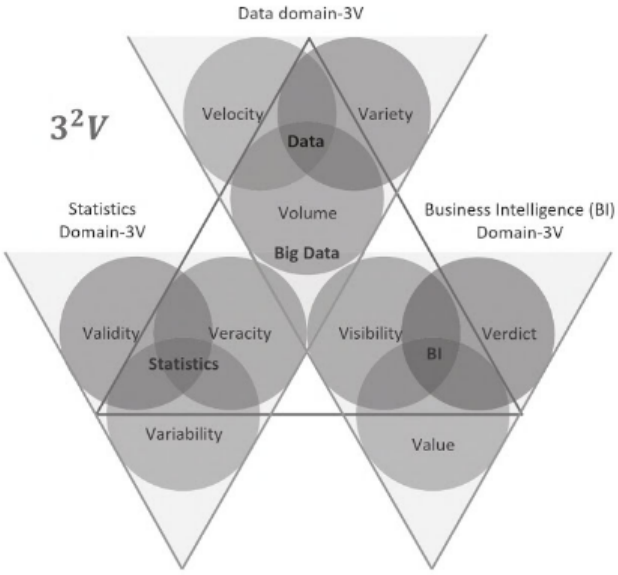

Figura 3.1: Diagrama de Venn de las 3^2 V's de Big Data.

En esta figura, se observa cómo los autores involucran nueve características de Big Data agrupadas en tres dominios del conocimiento:

- El dominio de los datos que se relaciona con la búsqueda de patrones donde se tiene en cuenta la velocidad, variedad y volumen; dominio en el cual el atributo crucial es el volumen ya que históricamente ha mostrado tener mayor variación respecto a las otras dos características.
- El dominio de la inteligencia de negocios en el cual se hacen predicciones y se integra el valor, la visibilidad y el veredicto como las características necesarias para tal fin, donde la visibilidad es el atributo clave dado que éste es el que permite obtener la predicción o la información en tiempo real a partir de ejercicios de Big Data, y finalmente,
- El dominio estadístico en el cual se realizan asunciones, donde la validez, veracidad y variabilidad deberían establecer los modelos estadísticos basados en las hipótesis correctas a partir de datos confiables y fuentes de datos fidedignas. En este dominio el factor trascendental es la veracidad, puesto que hace énfasis en cómo construir un modelo estadístico muy cercano a la realidad.

Además de esto, se puede notar que cada diagrama de Venn es soportado por una V en forma de triángulo que ilustra los atributos de las 3V's en un aspecto; así mismo, las tres características clave de cada V forman un único diagrama triangular jerárquico. Esto representa para los autores el significado principal de Big Data, ahora bien, la conexión de las 9V's en los tres dominios formando ese triángulo central, es lo que para ellos representa el sentido semántico del término, es decir, la relación entre los datos, la inteligencia de negocios y la estadística.

Para finalizar pese a que son variadas las opiniones en la literatura actual sobre lo que significa Big Data, en el contexto del análisis de datos en el que se trabajará, entendemos Big Data como la posibilidad de tomar mejores decisiones a partir de los ïnsights"derivados del procesamiento y análisis de ingentes y variadas cantidades de datos de los cuales se debe garantizar su confiabilidad, entendiéndose un insight como información que 1) tiene un entendimiento profundo, 2) es realmente significativa, 3) no es obvia y 4) es accionable, es decir, que se puede actuar en consecuencia a su obtención.

3.3. Características de Big Data

A partir de las tres V's propuestas por Doug Laney mostradas en la figura No. 3.2, han surgido varias características a tener en cuenta cuando se habla de Big Data como son volumen, velocidad, variedad, validez, veracidad, variabilidad, valor, visibilidad y veredicto. Ahora, examinaremos brevemente cada uno de estos atributos.

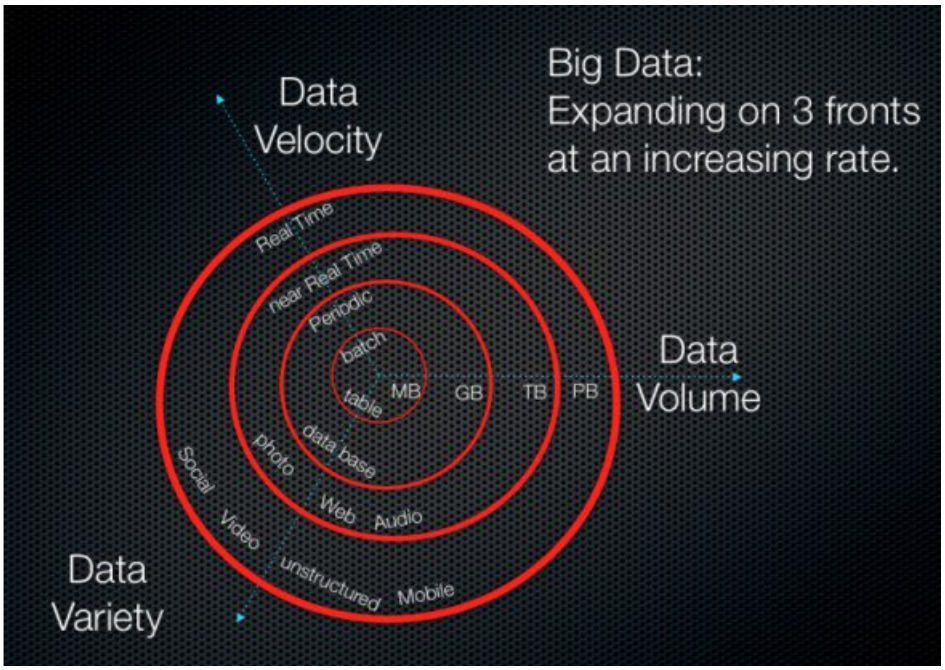

Figura 3.2: Las 3V's de Big Data, fuente [31]

Volumen

Es la cantidad de datos que tenemos - lo que solía ser medido en Gigabytes ahora se mide en Zettabytes (ZB) o incluso en Yottabytes (YB). El IoT (Internet de las cosas) es la fuente del crecimiento exponencial de los datos. Ahora que éstos se generan por las máquinas, las redes y la interacción humana en sistemas como medios de comunicación social, el volumen de datos a analizar es enorme. La siguiente tabla muestra la conversión de tamaños de datos de forma más clara.

Nombre	Símbolo	Potencias Binarias
byte	b	2^0
Kbyte	KB	2^{10}
Megabyte	MB	2^{20}
Gigabyte	GB	2^{30}
Terabyte	TB	2^{40}
Petabyte	PB	2^{50}
Exabyte	EX	2^{60}
Zettabyte	ZB	2^{70}
Yottabyte	yB	2^{80}

Cuadro 3.1: Tabla de equivalencias en tamaños de datos

- Cada día, el mundo produce 2,5 trillones de bytes de datos. Es decir 2,3 billones de gigabytes [32].
- En 2020, habremos creado 40 zettabytes de datos, es decir 43 billones de gigabytes.
- En 2014 los usuarios de Facebook compartían alrededor de 2.5 millones de piezas de contenido y se cree que Google recibió mas de 4 millones de consultas por minuto [33].

Variedad

Se refiere a las muchas fuentes y tipos de datos estructurados y no estructurados que existen. Cubre las diversas formas que pueden tomar los datos, a partir de datos tabulares cuidadosamente estructurados, semiestructurados y no estructurados como imágenes, correos electrónicos, hojas de cálculo, conversaciones de medios sociales y los medios de transmisión.

Para lograr entender a lo que se refiere específicamente los tres tipos de datos mencionados anteriormente, se explicará de la siguiente forma:

- **Datos Estructurados:** Son aquellos datos que son fáciles de almacenar, es decir que pueden ser almacenados en tablas con filas y columnas; además tienen una llave principal con la que pueden ser consultados fácilmente, como por ejemplo, en las bases de datos relacionales o en las hojas de cálculo.
- **Datos Semi-Estructurados:** Son aquellos datos que no son almacenados en bases de datos relacionales, pero cuentan con una organización interna con la cual es más sencilla de analizarlos, tales como documentos XML y bases de datos NoSQL. Es un tipo de datos estructurados, pero carece de la estricta estructura del modelo de datos. Con los datos semi-estructurados, etiquetas u otros tipos de marcadores se utilizan para identificar ciertos elementos dentro de los datos, pero los datos no tienen una estructura rígida.
- **Datos No Estructurados:** Son aquellos datos que no tienen un esquema o estructura totalmente alineada a un modelo de datos predefinido, por lo tanto, no pueden ser almacenados en una base de datos tradicional o predefinida. Los datos no estructurados son aquellos que no pueden ser fácilmente clasificados y no caben en una caja ordenada: fotografías e imágenes, videos, streaming de datos de instrumentos, páginas web, archivos PDF, presentaciones de PowerPoint, correos electrónicos, blogs, wikis, documentos de texto, por ejemplo.

Velocidad

Hace referencia a la velocidad en que los datos son accesibles. Se produce cuando los datos se pueden utilizar en tiempo real y necesitan ser tratados y almacenados de forma rápida. Un ejemplo es la elaboración de perfiles en tiempo real de visualización de anuncios de Internet que se adaptan de acuerdo a su patrón de uso. La velocidad refleja la frecuencia con la que los datos se generan, almacenan y comparten. Recientes investigaciones indican que no sólo los consumidores, sino también las empresas, generan más datos en menos tiempo.

Veracidad

Habla sobre la incertidumbre de los datos, es decir, la veracidad dice el nivel de confiabilidad asociada a ciertos tipos de datos. Es por eso, que se requiere un esfuerzo por obtener datos de alta calidad, teniendo presente que aunque se apliquen técnicas de limpieza, no se puede eliminar al 100 % la imprevisibilidad inherente de algunos datos como el tiempo, la economía o las futuras decisiones de compra que podrá tener un cliente. Se hace necesario entonces, reconocer y planificar dicha incertidumbre [3].

Variabilidad

Los flujos de datos, tanto en volumen como en variedad, pueden cambiar enormemente, pudiendo seguir un comportamiento cíclico predecible o siendo completamente aleatorio. Esta variabilidad es especialmente difícil de gestionar por la existencia de las redes sociales. Significa que los sistemas Big Data deben disponer del mismo tipo de elasticidad que es requerido en Cloud Computing y otros entornos virtualizados.

Visualización

Es el modo en el que los datos son presentados. Una vez que los datos son procesados (los datos están en tablas y hojas de cálculo), necesitamos representarlos visualmente de manera que sean legibles y accesibles, para encontrar patrones y claves ocultas en el tema a investigar. Existen herramientas de visualización que ayudan a comprender los datos gráficamente y en una perspectiva contextual. El uso de tablas y gráficos para visualizar grandes cantidades de datos complejos es mucho más eficaz para transmitir significado de hojas de cálculo e informes repletos de números y fórmulas.

Valor

Se obtiene valor de datos que se transforman en información; esta a su vez se convierte en conocimiento, y este en acción o en decisión. El valor de los datos está en que sean accionables, es decir, que los responsables de la empresas puedan tomar una decisión (la mejor decisión) en base a estos datos. La importancia, el valor o la utilidad de los datos es probablemente el atributo más relevante para las organizaciones. Los datos en sí mismos no tienen ningún valor.

Aunque el volumen, la velocidad y la variedad son intrínsecos al propio Big Data, las demás Vs mencionadas son atributos importantes que reflejan la complejidad gigantesca que presentan los grandes volúmenes de datos a la hora de procesarlos, analizarlos y beneficiarse de ellos.

Validez

Es la verificación de la información con la finalidad de saber si los datos son correctos y exactos para el uso previsto. Este atributo permite verificar la calidad de datos que son lógicamente sólidos; también es el proceso de inferencia basado en un modelo estadístico.

Veredicto

Es una elección o decisión potencial tomada por un administrador o comité basado en el alcance del problema, los recursos disponibles y cierta capacidad computacional. Esta es la V de mayor reto para ser cuantificada al inicio de Big Data. Si existen muchas hipótesis o "Que pasa si", el costo de colectar, recuperar datos y ETL (extraer, transformar y cargar), especialmente para extraer datos archivados, será costoso [9, Pag. 12].

3.4. ¿Qué es la Datificación?

Se ha usado de manera indiscriminada los términos "datos" e "información" pero antes de entrar a explicar la relación que existe entre la datificación y Big Data, vamos a diferenciarlos.

Se entiende por información la colección general de conocimiento relevante sobre algo en particular. Es sinónimo al uso coloquial que se le da al término dato, pero entiéndase que éste último se refiere a la información que no puede ser derivada de cualquier otra cosa, es decir, que sirve como el axioma a partir del cual se deriva todo lo demás. Un ejemplo de dato puede ser una fecha de nacimiento ya que en base a ello, se puede calcular la edad de la persona.

Habiendo aclarado la relación entre ambos términos, se encontró que la datificación es "el proceso por el que se plasma un fenómeno (incluso un estado de ánimo) en un formato cuantificado para su tabulación y análisis" [34]. Actualmente vivimos en un entorno donde cada vez más se pretende producir datos que puedan ser leídos y medidos a través de una infraestructura tecnológica; el mundo entero está siendo datificado y todo este esfuerzo ha dado como resultado al fenómeno "Big Data".

4

Análisis Big Data

En el capítulo anterior se dió una vista general al concepto de Big Data y sus atributos, ahora vamos a enfocarnos en lo que hace referencia al análisis Big Data para posteriormente observar el impacto que tiene la aplicación de este tipo de técnicas sobre el estudio de datos hidroclimatológicos.

Teniendo en cuenta las características mencionadas que describen el término, el análisis Big Data corresponde a la aplicación de técnicas de análisis avanzadas y en tiempo real sobre conjuntos de datos que poseen esos atributos. En el caso de IBM se considera que se debe emplear este tipo de técnicas sobre datos que tengan una o varias de las tres Vs principales: gran volumen, variedad o alta velocidad, pero en general, las organizaciones conciben el análisis Big Data sobre "grandes cantidades o volumenes de datos", con el fin de tomar mejores decisiones de manera rápida y oportuna a partir de datos que eran inaccesibles o inutilizados.

Según SAS [35] el análisis Big Data examina ingentes cantidades de datos para descubrir patrones ocultos, correlaciones y otros "insights". Con la tecnología de hoy, es posible analizar los datos y obtener respuestas a partir de ellos casi de forma inmediata, hecho que no sucede con soluciones de inteligencia de negocios más tradicionales. Es allí donde el análisis Big Data cobra relevancia, dado que ayuda a las organizaciones a aprovechar sus datos y a usarlos para identificar nuevas oportunidades.

El hecho de que las organizaciones posean la habilidad de trabajar con mayor rapidez y agilidad, hace que Big Data se convierta en una herramienta fundamental: operaciones más eficientes, reducción de costos, predicciones con un mayor nivel de exactitud, creación de nuevos productos o servicios, fidelización de los clientes, entre otras ventajas que cada empresa puede hallar al aplicar este tipo de soluciones en TI.

Los análisis Big Data son diseñados para explorar y aprovechar las características únicas de los datos desde minería secuencial/temporal y minería espacial, hasta minería para flujos de datos de alta velocidad y datos de sensores [1, Pag. 6].

El análisis es formulado en base a potentes técnicas matematicas incluyendo aprendizaje de máquina (Machine Learning), redes bayesianas, modelos de Markov ocultos, modelos de aprendizaje supervisado (Support Vector Machines), aprendizaje de refuerzo (Reinforcement Learnig) y modelos de conjunto. Además se cuenta con otras técnicas como la minería de datos, el aprendizaje profundo (Deep Learning) y el análisis de texto que se ayuda de las técnicas de procesamiento de lenguaje natural y se orienta en la recuperación de la información y la lingüística computacional, para realizar actividades como detección de eventos, seguimiento de tendencias, análisis de sentimientos, modelado de temas y minería de opinión. A su vez hoy se habla de técnicas como el análisis de video, de audio, de redes sociales, análisis web, entre otras.

El proceso general de extraer información de Big Data puede dividirse en cinco etapas [36] como se muestra en la figura 4.1:

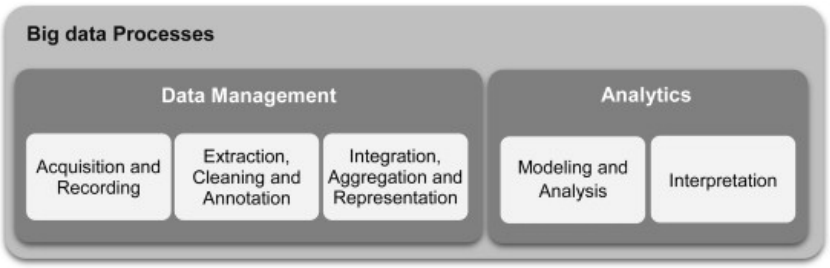

Figura 4.1: Etapas de extracción de la información

El proceso Big Data a su vez se fracciona en dos subprocesos principales: la administración de datos y el análisis. El primero, involucra procesos y tecnologías de apoyo para adquirir y almacenar datos, prepararlos y recuperarlos para su análisis. El segundo, por otra parte, se refiere a las técnicas usadas para analizar y obtener inteligencia desde Big Data, es por esto que el análisis Big Data se puede ver como un subproceso del proceso general de **extraer insigths** de los datos.

Existen varios aspectos que se deben apreciar al momento de sumergirse en el análisis Big Data [37]:

- **El descubrimiento:** En muchos casos no se conoce realmente qué es lo que se tiene y cómo los conjuntos de datos se relacionan entre sí, es por tal motivo que se deben realizar procesos de exploración y descubrimiento.

- **La iteración:** Las relaciones actuales entre los datos no siempre se conocen de antemano, descubrir *insight* es con frecuencia un proceso *iterativo* al encontrar las respuestas buscadas. La naturaleza de la iteración es que a veces nos lleva por un camino que resulta ser un callejón sin salida y no está mal, dado que la experimentación es parte del proceso de aprendizaje; los expertos en análisis sugieren que se debe empezar con proyectos pequeños y bien definidos, aprender de cada iteración y gradualmente moverse a la siguiente idea o campo de acción.
- **Capacidad Flexible:** La naturaleza iterativa del análisis Big Data requiere la inversión de más tiempo y recursos para resolver problemas.
- **Minería y Predicción:** A medida que se explora los datos para descubrir patrones y relaciones, el análisis predictivo puede producir los *insights* buscados, ya que como se mencionaba anteriormente, no siempre se conoce cómo los diferentes elementos de los datos se relacionan entre sí.
- **Gestión de Decisiones:** Es necesario considerar la velocidad y volumen de la transacción, por ejemplo, si lo que se busca es personalizar un sitio web, se requiere tener en cuenta como se automatizará y optimizará la implementeación de las acciones que permitirán conseguir el objetivo.

En la exploración y busqueda de las relaciones entre los datos no solo se debe saber el qué, sino también el porqué. Es allí donde toman gran participación las técnicas que se describirán a continuación: se desea por ejemplo entender por qué unos pacientes sobreviven más que otros, por qué hay tráfico en las carreteras de una ciudad en un horario determinado, por qué tienen tan alto impacto los desbordamientos de los ríos a tal punto de ocasionar la muerte de tantas personas, etc.

Sin duda alguna, en el campo de investigación de este proyecto, saber el qué y el por qué marca un hito de gran relevancia en el análisis de datos hidroclimatológicos: permite tomar mejores decisiones en torno a la prevención de desastres en nuestro hermoso país y no solo en ese aspecto, sino en cómo las personas (gobierno, afectados, observadores) reaccionan frente a trágicos acontecimientos como el que ocurrió el pasado 1 de Abril en Mocoa la capital de Putumayo. El análisis Big Data le apuesta a dejar de lado "lo que yo creo o pienso" y reemplazarlo por la toma conciente de decisiones basadas en información coherente y sustentable derivada de datos reales.

4.1. Tipos de Análisis

4.1.1. Análisis Prescriptivo

Este tipo de análisis revela qué acciones deberían ser tomadas y para ello se utilizan algoritmos de optimización y simulación. Es el tipo de análisis más valioso y usualmente resulta en reglas y recomendaciones para posteriores pasos. Aunque es realmente valioso no es el más empleado (dado que es relativamente complejo de administrar), pues según Gartner [38], el 13 % de las organizaciones están usando la predicción y sólo el 3 % el análisis prescriptivo.

La analítica prescriptiva provee un enfoque "láser" para dar respuesta a una pregunta específica, ayuda a determinar la mejor solución entre una variedad de opciones dados los parámetros conocidos y sugiere opciones para aprovechar una oportunidad futura o mitigar un futuro riesgo. Intenta cuantificar el efecto de las decisiones a tomar y predice no sólo lo que sucederá, sino también por qué ocurrirá proporcionando recomendaciones sobre acciones que tomarán ventaja de las predicciones. Por ejemplo en la industria de la salud, la analítica prescriptiva permite medir el número de pacientes que son clínicamente obesos y a continucación agregar filtros para factores como la diabetes y los niveles de colesterol para determinar dónde enfocar el tratamiento. Este tipo de análisis se puede aplicar en general en cualquier industria.

4.1.2. Análisis Predictivo

El análisis predectivo tiene su raíz en la habilidad de "Predecir" lo que sucederá, busca entender el futuro y provee "insights" accionables basados en los datos. Los análisis predictivos proporcionan estimaciones sobre la probabilidad de un resultado futuro; es importante recordar que ningún algoritmo estadístico puede "predecir" el futuro con el 100 % de certeza. Dichas estadísticas tratan de tomar los datos que se tienen y completar los datos faltantes con las mejores conjeturas. Combinan datos históricos para identificar patrones y aplicar modelos y algoritmos estadísticos para capturar relaciones entre varios conjuntos de datos.

4.1.3. Análisis Diagnóstico

El análisis diagnóstico es empleado para determinar por qué sucedió algo y es caracterizado por técnicas tales como el descubrimiento de datos, la minería de datos y las correlaciones. Tiene una mirada más profunda a los datos para intentar comprender las causas de los eventos y comportamientos.

4.1.4. Análisis Descriptivo

El análisis descriptivo hace exactamente lo que su nombre indica, "describir" o resumir datos sin procesar (crudos) y hacerlos entendibles e interpretables por los seres humanos. Son análisis que describen el pasado, entendiéndose por pasado, cualquier punto del tiempo en el que un evento se ha producido, ya sea hace un minuto, o hace años [39]. La analítica descriptiva es útil porque permite aprender de los comportamientos pasados y entender cómo pueden influir en los resultados futuros. La gran mayoría de la estadística que implementamos entra en esta categoría, por ejemplo si se piensa en la aritmética básica como sumas, promedios, cambios porcentuales. Los datos subyacentes de este tipo de análisis son un recuento o agregación de una columna de datos filtrada a la que se aplica matemáticas básicas. Ejemplos comunes de la analítica descriptiva son los informes que proporcionan información histórica sobre la producción de una empresa, las finanzas, operaciones, ventas, etc.

4.2. Técnicas Utilizadas en el Análisis Big Data

En esta sección abordaremos las técnicas de análisis que representan un subconjunto importante, puesto que en la actualidad se posee un amplio numero de ellas y abordarlas todas no es el objetivo de nuestra investigación. Empezaremos describiendo las técnicas más tradicionales como *Machine Learning* hasta las más nombradas en la actualidad como lo es el *análisis de texto*.

4.2.1. Data Mining

La minería de datos es la ciencia dedicada a obtener información desde los datos y conocimiento a partir de la información. Este campo provee la teoría y metodologías sobre cómo y cuánta información útil puede ser obtenida de los conjuntos de datos y cómo adquirir conocimiento de dicha información. Gracias al advenimiento de Big Data, las técnicas de la minería de datos han cobrado mayor importancia, técnicas que permiten separar datos útiles de improductivos, detectar ruido en los datos, modelar el sistema generando tales datos, hacer muestreo de los datos mientras se preservan sus propiedades, y aún más.

4.2.2. Machine Learning

La esencia de ML es un proceso automático de reconocimiento de patrones mediante un aprendizaje de máquina. El objetivo principal de ML es construir sistemas que puedan desarrollar en el mismo o mayor nivel la competencia humana para el manejo de muchas tareas complejas o problemas. A su vez, hace parte de la inteligencia artificial y sus aplicaciones incluyen la toma de decisiones, la predicción y es una tecnología clave que permite la implementación de la minería de datos y las tecnicas Big Data en campos diversos como la salud, la ciencia, la ingeniería, los negocios y las finanzas.

Inicialmente, la meta de Machine Learning era construir robots y simular las actividades humanas, después la aplicación de la inteligencia artificial estaba siendo generalizada para resolver problemas generales por una máquina. La solución popular era alimentar un computador con algoritmos (o una secuencia de instrucciones) para que éste pudiera transformar los datos de entrada en la respuesta de salida. Esto era llamado con frecuencia un sistema basado en reglas o sistema experto. Sin embargo, no se puede encontrar facilmente algoritmos adecuados para muchos problemas, por ejemplo, para el reconocimiento de la escritura humana. No se sabe cómo transformar la entrada de una carta escrita a mano por una salida de una carta estándar reconocida; una alternativa es aprender de los datos y ese principio de aprendizaje es similar al de "ensayo y error" o "La sabiduría de la multitud" [9, Pág. 15]. Esto quiere decir que teniendo solo una prueba, se incurriría en un alto margen de error, pero si se incluyen muchas pruebas, el error reducirá a un nivel de tolerancia aceptable o convergente.

En orden para descubrir patrones y tendencias significativos, desde un enorme conjunto de datos, la estadística es la herramienta fundamental para añadir valor al muestreo, modelamiento, análisis, interpretación y presentación de los datos.
Los componentes principales en la definición de Machine Learning son:

- El entrenamiento de la máquina para que aprenda automáticamente y mejore los resultados en tanto que obtenga más datos.
- Descubrir o reconocer patrones e inteligencia con los datos de entrada.
- Hacer predicción sobre entradas de datos desconocidas.
- La máquina irá adquiriendo conocimiento directamente de los datos y resolverá problemas.

Hablando ampliamente, las tareas de Machine Learning pueden ser categorizadas en los tipos que se describirán a continuación: el aprendizaje supervisado y el aprendizaje no supervisado [40].

4.2.2.1. Aprendizaje Supervisado

La tarea de aprendizaje consiste en generalizar desde un conjunto de entrenamiento, que es etiquetado por un "supervisor"para contener información sobre la clase de una muestra, de modo que se puedan hacer predicciones sobre nuevas muestras, aunque éstas no hayan sido "vistas" por el algoritmo entrenado. Si la salida (o predicción) pertenece a un conjunto continuo de valores, entonces tal problema se denomina **regresión**, mientras que si la salida asume valores discretos, entonces el problema se llama **clasificación**.

Con el aprendizaje supervisado, la salida del algoritmo ya se conoce -al igual que cuando un estudiante está aprendiendo de un instructor-. Todo lo que hay que hacer es elaborar el proceso necesario para obtener de su entrada, la salida (mapeo). A continuación presentamos brevemente algunas técnicas de clasificación.

Support vector machines:

Es una técnica de aprendizaje supervisado ampliamente utilizada que es notable por ser práctica y a su vez teóricamente sólida [41]. La máquina vectorial de soporte (SVM) es un enfoque estadístico que permite la predicción no lineal del tipo que los análisis tradicionales no permiten. Imagine que la SVM se utiliza para clasificar los casos en uno de los dos valores de una variable de criterio (por ejemplo, los vendedores que se quedan frente a la salida voluntaria de la empresa), basado en un conjunto de predictores (por ejemplo, satisfacción laboral de los vendedores, desempeño y salario actual). SVM considerará que cada persona en la muestra tiene un vector de p predictores; entonces tratará de separar los vectores que pertenecen a cada valor de criterio tan limpio como sea posible utilizando lo que se llama un hiperplano.

La cantidad de separación se denomina margen, donde un margen más amplio es mejor, siempre y cuando se optimicen los errores en la clasificación. Los vectores de soporte son los datos que se encuentran a lo largo del margen y sirven para definirlo. El acercamiento a SVM usando márgenes suaves permite que la solución final tenga algunos errores en la clasificación. Esto en realidad podría tener que hacerse en muchos casos, porque la separación no puede hacerse de manera totalmente limpia. El enfoque SVM puede extenderse a más de dos categorías; puede modificarse para reducir la influencia negativa de los valores

atípicos y puede aplicarse a datos que experimentan una transformación no lineal (por ejemplo, un núcleo polinomial) con la esperanza de mejorar la separación y por lo tanto la eficiencia de clasificación.

Artificial neural networks:

Las redes neuronales artificiales son similares a la regresión múltiple, donde se determinan las relaciones entre los múltiples predictores y un criterio único. Sin embargo, en lugar de utilizar un modelo lineal, el modelo de ANN es más complejo, puesto que las relaciones entre predictores y criterios se basan en un modelo cuyos componentes se asemejan a las neuronas en el cerebro.

Más específicamente, las neuronas son nodos ocultos entre los predictores y el criterio, donde el usuario tiene mucha latitud para determinar el número de neuronas, cómo están conectadas y cuántas capas de conexión hay. En este sentido, la ANN es un método de "caja negra", ya que aunque se observan los predictores y el criterio, las neuronas (nodos) están ocultas y normalmente no hay una sola manera correcta de especificar su número o configuración.

Suponiendo que el usuario especifica una ANN para propósitos de predicción; entonces cada neurona recibe la activación, que es alguna función de la actividad entrante (generalmente una suma ponderada) de las neuronas conectadas. Una vez que la activación recibida excede un determinado umbral, la neurona entonces 'dispara' o transmite su propia activación a otras neuronas conectadas en la red neuronal que funcionan de manera similar. Normalmente, las neuronas sólo transmiten a través de capas neuronales, aunque algunos modelos permiten también la transmisión hacia atrás (backpropagation). Las ANN utilizan métodos iterativos para lograr el conjunto apropiado de pesos en la red neuronal; al hacerlo, también a menudo incorporan propagación hacia atrás, parte del procedimiento de iteración en ANN que examina discrepancias entre los valores de criterio predichos y reales, revisa la red y ajusta los pesos neurales para reducir esas discrepancias antes de iniciar la siguiente iteración de predicción por el modelo.

Una ventaja es que aun cuando se tienen pocas neuronas conectadas en la red, ésta permitirá realizar predicciones complejas; para un fenómeno más complejo se requerirán más neuronas, sin embargo, en ese caso se podría llegar a tener un modelo sobre-entrenado que no es conveniente para llevar a cabo predicciones en nuevos conjuntos de datos. Algo interesante es que los modelos de redes neuronales han demostrado poseer la misma capacidad de predicción que los modelos de regresión lineal para la predicción de la personalidad en

el desempeño laboral de grandes muestras de profesionales, reclutas policiales y conductores de autobuses. En tanto que los modelos de regresión contenían una importante interacción y en términos cuadrados, estos modelos simples funcionan bien (y a veces mejor) si se configuran como redes neuronales.

Con el arrivo de Big Data y el uso de técnicas de modelado flexible como las redes neuronales artificiales, los investigadores pueden descubrir inductivamente interacciones no lineales y confiables con mayor facilidad, para lo cual se requiere tener en cuenta que una red neuronal basada en la cantidad y la conexión de las neuronas puede revelar diferentes no linealidades para el mismo conjunto de datos, por lo tanto es necesario que dichas redes sean sometidas a altos niveles de escrutinio, tales como la prueba con modelos alternativos, validación cruzada sobre muestras independientes de datos, entre otros.

Decision trees & random forests:

Los árboles de decisión (Decision trees) son un modelo que usa un conjunto de reglas binarias aplicadas para calcular un valor objetivo. Pueden ser usados para aplicaciones de clasificación (variables categóricas) o regresión (variables continuas). La idea en este tipo de modelos es separar los datos (node padre) en dos subconjuntos (nodos hijos) a través del cálculo de la mejor característica divisora determinada por un criterio de partición elegido, entonces los dos conjuntos resultantes llegan a ser nuevos nodos padres y son subsecuentemente divididos en dos nuevos nodos hijos. La división binaria continúa hasta que todas las observaciones son clasificadas [42].

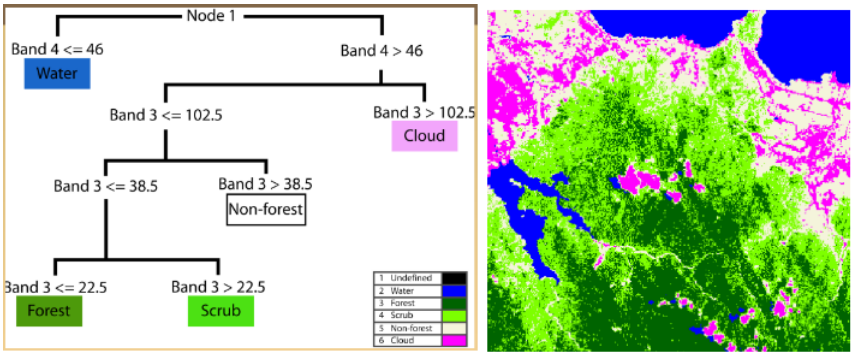

Figura 4.2: Ejemplo de clasificación usando árboles de decisión, fuente [43]

Random Forest (RF) es un clasificador de conjunto que utiliza muchos modelos de árboles de decisión [43]. Un modelo de conjunto combina los

resultados de diferentes modelos que pueden ser de un tipo similar o diferente, donde el resultado de un modelo de conjunto es tipicamente mejor que el de uno de los modelos individuales. La información de la presición y la variable de importancia es provista con el resultado.

El algoritmo random forest funciona de la siguiente manera: cada vez que se suple una entrada para el RF, ésta se hace pasar bajo cada uno de los árboles de decisión que constituyen el bosque. Cada árbol predice independientemente una clasificación y "vota" por la clase correspondiente. La mayoría de votos decide el RF total. Este voto agregado de varios árboles posee intrínsicamente menos ruido y suceptibilidad ante los valores atípicos comparado a la salida de un único árbol de decisión.

Naïve Bayes:

Es un método que puede utilizarse en una variedad de configuraciones, pero es el que más se usa en la clasificación de texto basado en palabras que han sido extraídas del mismo. Naïve Bayes ha sido aplicado como una herramienta para la detección y filtrado de mensajes spam en los servidores de correo. También puede ser usado en investigaciones organizacionales para clasificar si las personas deciden o no aplicar a un empleo basados sobre información detallada que adquieren sobre la organización, el empleo y los beneficios ofrecidos [41].

4.2.2.2. Aprendizaje No Supervisado

Se caracteriza porque no posee un conjunto de datos para el entrenamiento del modelo y además las salidas son desconocidas. El objetivo del aprendizaje no supervisado es modelar la estructura o distribución en los datos con el fin de aprender más sobre ellos.

Los problemas de aprendizaje no supervisado pueden ser agrupados en clustering o asociación. El primero se refiere al descubrimiento de los grupos inherentes en los datos, tales como la agrupación de clientes mediante el análisis del comportamiento. La asociación por su parte busca descubrir reglas que describen grandes porciones de los datos, como por ejemplo, las personas que compran el producto X, también tienden a llevar el producto "y".

Ejemplos de algoritmos de aprendizaje no supervisado son k-means y aprendizaje de reglas de asociación.

K-means:

Es un algoritmo de clustering que se aplica sobre datos no categorizados con el objetivo de encontrar grupos de datos, donde el número de grupos es representado por la variable **K** [44]. El algoritmo trabaja iterativamente para asignar cada punto de datos a uno de los grupos K basados en las características que se proporcionan. Los puntos de datos se agrupan en función de la similitud de características. Los resultados del algoritmo de agrupamiento de K-means son:

1. Los centroides de los K clusters, que pueden ser utilizados para clasificar nuevos datos.

2. Las clasificaciones para los datos de entrenamiento (cada punto de datos es asignado a un solo cluster).

Cada centroide de un cluster es una colección de valores característicos que definen los grupos resultantes. Examinar los pesos de la característica del centroide permite interpretar cualtitativamente el tipo de clúster que cada grupo representa.

4.2.3. Reinforcement Learning

El aprendizaje por refuerzo (RL) es una rama de Machine Learning que permite a las máquinas y a los agentes de software determinar automáticamente el comportamiento ideal sin un contexto específico, en orden de mejorar su desempeño. Se tiene el concepto de recompensa y se requiere retroalimentación simple de ésta para que el agente aprenda su comportamiento, esto es entendido como la **señal de refuerzo**. Se supone que en un problema el agente toma la mejor decision a seleccionar basándose en su estado actual; cuando dicho proceso se repite, el problema es conocido como el *Proceso de Decisión de Markov*.

4.2.4. Deep Learning

El aprendizaje profundo es una técnica para la implementación de Machine Learning. Es el uso de redes neuronales artificiales que contienen más de una capa oculta en forma de estructuras lógicas que se asemejan en mayor medida a la organización del sistema nervioso de los mamíferos, teniendo capas de unidades de proceso o neuronas artificiales que se especializan en detectar determinadas características en los objetos percibidos. Esas redes emplean un esquema de cascada en el cual cada nivel usa la salida del anterior como su entrada. La red

aprende múltiples categorías de representación que corresponden a los diferentes niveles de abstracción, formando de esta manera una arquitectura de conceptos. Deep learning le permite a una máquina aprender conceptos complejos a partir de conceptos simples.

4.2.5. Data Visualization

La visualización de datos es en general un término que describe algún esfuerzo por ayudar a las personas a entender el significado de los datos a través de la ubicación de los mismos en un contexto visual. La visualización de datos efectiva es clave para obtener valor de una inversión en el análisis de exhuberantes cantidades de datos y es por ello que según Mc Daniel, cofundador de Freakalaytics LLC [45], hay algunas limitaciones que se deben considerar antes de sumergirse en el proceso de visualizar los resultados del análisis Big Data.

La primera limitación es encontrar la combinación de datos correcta para las herramientas de visualización de datos, puesto que aún si se tuvieran monitores avanzados no sería posible mostrar billones de puntos de datos individuales en la publicación de un grafo, dada la capacidad de pixeles que se pueden mostrar en la pantalla. Es allí donde las organizaciones deben elegir que tipo de datos agregar o eliminar en la visualización valiéndose de los requisitos específicos del público objetivo de la visualización y de técnicas de limpieza de datos que permitan dar confiabilidad en los resultados. Además se deben elegir las gráficas adecuadas para mostrar dichos resultados.

La segunda limitación se refiere al problema de la visualización de datos: "mucha información". La solución es entonces enfocarse en las preguntas que necesitan respuestas y mantener la visualización tan clara y limpia como sea posible. La utilización de gráficos en 3D si bien es muy llamativa e interesante, se debe manejar con cuidado, dado que hay evidencia de que nuestra mente no puede procesarlos y entenderlos a la perfección. Para aumentar la efectividad en la representación de los datos se requiere incrementar la comunicación entre las partes interesadas: por ejemplo entre los gerentes y el departamento de Tecnologías de la información de la compañía.

4.2.6. Text Analytics

El análisis de texto o **minería de texto** es el proceso de analizar texto sin una organización formal o estructura para extraer información relevante. Algunos

ejemplos de datos textuales usados por las organizaciones son retroalimentaciones de los usuarios en las redes sociales, correos, blogs, foros en línea, respuestas de encuestas, documentos corporativos, noticias y logs del centro de llamadas, entre otros. El proceso de análisis y levantamiento en el análisis de texto hace uso de métodos que involucran análisis estadístico, linguística computacional (NPL) y aprendizaje de máquina. Una de las técnicas es la extracción de la información (Information Extraction) implementada para extraer datos estructurados desde textos sin estructura, donde se ejecutan dos subtareas; la primera es el reconocimiento de entidades (encuentra nombres en el texto y los clasifica en categorías predefinidas como personas, ubicaciones y fechas) y la segunda es la extracción de relaciones (encuentra y extrae las relaciones semánticas entre las entidades) [46]. Los métodos usados para el reconocimiento de entidades son técnicas de aprendizaje supervisadas (modelos de Markov ocultos, modelo de entropía, SVM, campos aleatorios condicionales), enfoques de aprendizaje semi-supervisado (bootstrapping) y aprendizaje no supervisado (algoritmos de clustering).

Según Dan Sullivan experto en minería de datos [47], se cuenta con varias herramientas y técnicas para obtener valiosos *insights*:

1. Análisis de sentimientos: Analiza la opinión o la intención de los comentarios de las personas sobre un tema determinado en las redes sociales o a través de un centro de llamadas con el objetivo de entender las necesidades de los usuarios para brindarles mejores soluciones. Las siguientes son formas de analizar los sentimientos:

   Análisis de polaridad: Se identifica el "tono" de las comunicaciones (positivas o negativas).

   Categorización: Las herrramientas se hacen más finas, es decir, se determina si alguien está confundido o enojado, por ejemplo.

   Ubicar una escala sobre la emoción: de "triste" a "feliz" y de 0-10, por ejemplo.

2. Modelado de tópicos: Es una técnica de gran usabilidad para identificar temas dominantes en una vasta colección de documentos.

3. Frecuencia de término: Examina la frecuencia con la que una palabra aparece en un documento y su importancia relativa respecto a todo el conjunto de documentos analizados. Esto puede ser aplicado para construir clasificadores o modelos predictivos.

4. Reconocimiento de entidad nombrada: Básicamente busca reconocer sustantivos y podría ser usado para extraer personas, organizaciones, ubicaciones geográficas, fechas o similares en el texto.

5. Extracción de eventos: Va más allá que la técnica anterior, puesto que no sólo busca sustantivos sino también las relaciones entre ellos y el tipo de inferencias que se pueden obtener desde los sucesos referidos en el texto.

Un ejemplo de software para el análisis de texto es "SPSS Text Analytics" de IBM, donde hacen la promesa de convertir datos no estructurados de encuestas, en datos cuantitativos para extraer conocimiento mediante la aplicación del análisis de sentimientos y de tecnologías como el procesamiento de lenguaje natural (NLP) [48].

4.3. Herramientas y tecnologías Big Data

Para implementar técnicas de análisis avanzadas es necesario contar con sistemas de análisis de alto desempeño, es por tal motivo que en la actualidad se cuenta con tecnologías como Hadoop, Spark, Storm, Hive, Pig, MapReduce, entre otras. A continuación se hará mención a éstas.

MapReduce

MapReduce es un framework que proporciona un sistema de procesamiento de datos paralelo y distribuido. Su nombre se debe a las funciones principales que son Map y Reduce. MapReduce está pensado para la solución práctica de algunos problemas que pueden ser paralelizados, pero se ha de tener en cuenta que no todos los problemas pueden resolverse eficientemente con MapReduce. MapReduce está orientado a resolver problemas con conjuntos de datos de gran tamaño, por lo que utiliza el sistema de archivos distribuido HDFS [49]. Las funciones Map y Reduce están definidas ambas con respecto a datos estructurados en tuplas del tipo (clave, valor).

Función Map()

Map toma uno de estos pares de datos con un tipo en un dominio de datos, y devuelve una lista de pares en un dominio diferente:

$$\text{Map(k1,v1)} \rightarrow \text{list(k2,v2)}$$

La función map se encarga del mapeo y es aplicada en paralelo para cada ítem en la entrada de datos. Esto produce una lista de pares (k2,v2) por cada llamada. Después de eso, el framework de MapReduce junta todos los pares con la misma clave de todas las listas y los agrupa, creando un grupo por cada una de las diferentes claves generadas. Desde el punto de vista arquitectural el nodo master toma el *input*, lo divide en pequeñas piezas o problemas de menor identidad, y los distribuye a los denominados *worker nodes*. Un nodo trabajador puede volver a sub-dividir, dando lugar a una estructura arbórea,procesa el problema y pasa la respuesta al *nodo maestro*.

Función Reduce()

La función reduce es aplicada en paralelo para cada grupo, produciendo una colección de valores para cada dominio:

$$\text{Reduce(k2, list (v2))} \rightarrow \text{list(v3)}$$

Cada llamada a Reduce típicamente produce un valor v3 o una llamada vacía, aunque una llamada puede retornar más de un valor. El retorno de todas esas llamadas se recoge como la lista de resultado deseado.

Por lo tanto, el framework MapReduce transforma una lista de pares (clave, valor) en una lista de valores. Este comportamiento es diferente de la combinación *map and reduce* de programación funcional, que acepta una lista arbitraria de valores y devuelve un valor único que combina todos los valores devueltos por mapa.

Hadoop

Hadoop es un sistema de código abierto que se utiliza para almacenar, procesar y analizar grandes volúmenes de datos. Sus ventajas son muchas [50]:

- Aísla a los desarrolladores de todas las dificultades presentes en la programación paralela.
- Cuenta con un ecosistema que sirve de gran ayuda al usuario, ya que permite distribuir el fichero en nodos, que no son otra cosa que ordenadores con commodity-hardware.
- Es capaz de ejecutar procesos en paralelo en todo momento.
- Dispone de módulos de control para la monitorización de los datos.

- Presenta una opción que permite realizar consultas.
- También potencia la aparición de distintos add- ons, que facilitan el trabajo, manipulación y seguimiento de toda la información que en él se almacena.

Los componentes básicos de Hadoop son el sistema de archivo distribuido (HDFS), que permite que el fichero de datos no se guarde en una única máquina sino que sea capaz de distribuir la información a distintos dispositivos y el framework Mapreduce. La gran ventaja de Hadoop es que hace posible escoger y utilizar el lenguaje y las herramientar más adecuadas para la tarea concreta que se va a realizar.

Spark

Spark es una tecnología de código abierto (cluster-computing), originalmente desarrollada en la Universidad de California y más tarde fue donada a la *Fundación de Software Apache*, que la ha mantenido desde entonces. Spark proporciona una interfaz para la programación de Big Data con implícita paralelismo de datos y de tolerancia a fallos [51].

Spark es una plataforma de computación de código abierto para análisis y procesos avanzados, que tiene muchas ventajas sobre Hadoop. Desde el principio, Spark fue diseñado para soportar en memoria algoritmos iterativos que se pudiesen desarrollar sin escribir un conjunto de resultados cada vez que se procesaba un dato. Esta habilidad para mantener todo en memoria es una técnica de computación de alto rendimiento aplicado al análisis avanzado, la cual permite que Spark tenga unas velocidades de procesamiento que sean 100 veces más rápidas que las conseguidas utilizando MapReduce.

Spark tiene un framework integrado para implementar análisis avanzados que incluye la librería MLlib, el motor gráfico GraphX, Spark Streaming, y la herramienta de consulta Shark. Esta plataforma asegura a los usuarios la consistencia en los resultados a través de distintos tipos de análisis.

Hive

Apache Hive [52] es una infraestructura de almacenamiento de datos construida sobre Hadoop para proporcionar agrupación, consulta, y análisis de datos. Inicialmente desarrollado por Facebook, Apache Hive es ahora utilizada y desarrollada por otras empresas como Netflix y Financial Industry Regulatory

Authority (FINRA). Amazon mantiene una derivación de software de Apache Hive incluida en Amazon Elastic MapReduce en sus servicios web (AWS).

Apache Hive soporta el análisis de grandes conjuntos de datos almacenados bajo HDFS de Hadoop y en sistemas compatibles como el sistema de archivos Amazon. Ofrece un lenguaje de consultas basado en SQL llamado HiveQL, con esquemas para leer y convertir consultas de forma transparente en MapReduce, Apache Tez y tareas Spark. Los tres motores de ejecución pueden correr bajo YARN. Para acelerar las consultas, Hive provee índices, que incluyen índices de bitmaps.

Otras características de Hive incluyen:

- Indexación para proporcionar aceleración, tipo de índice que incluye compactación e índices de bitmaps. Otros tipos de índices serán incluidos en futuras versiones.
- Diferentes tipos de almacenamiento como texto, RCFile, HBase, ORC, y otros.
- Almacenamiento de metadatos en bases de datos relacionales, lo que permite reducir el tiempo para realizar verificaciones semánticas durante la ejecución de consultas.
- Operaciones sobre datos comprimidos almacenados en el ecosistema Hadoop usando algoritmos que incluyen DEFLATE, BWT, snappy, etc.
- Funciones definidas por el usuario (en inglés, user-defined function, UDF) para manipular fechas, textos, y otras herramientas de minería de datos. Hive soporta la extensión de las funciones definidas por el usuario de manera de tratar casos no contemplados.
- Consultas estilo SQL (HiveQL), las cuales son convertidas automáticamente a MapReduce o Tez, o tareas Spark.

Por defecto, Hive almacena sus metadatos en una base de datos apache Derby, pero puede ser configurado para usar MySQL.

Storm

Apache Storm es un marco de procesamiento de flujo que se centra en una latencia extremadamente baja y es quizás la mejor opción para cargas de trabajo que requieren un procesamiento casi en tiempo real [53]. Puede manejar grandes cantidades de datos y entregar resultados con menor latencia que otras soluciones.

El procesamiento de flujos de tormenta funciona mediante la orquestación de los DAG (Dirigido acíclicos gráficos) en un marco que llama topologías. Estas topologías describen las diversas transformaciones o pasos que serán tomadas en cada pieza de entrada de datos a medida que entra en el sistema.

"Storm es más que un sistema tradicional de analítica de big data: es un ejemplo de un sistema complejo de procesamiento de eventos (CEP). Los sistemas CEP son normalmente categorizados como orientados a la computación y a la detección, cada una de los cuales puede ser implementada en Storm mediante algoritmos definidos por el usuario. Los CEPs pueden, por ejemplo, utilizarse para identificar eventos significativos a partir de una gran cantidad de eventos y después actuar sobre esos eventos en tiempo real. Nathan Marz proporciona varios ejemplos en los que se usa Storm dentro de Twitter. Uno de los más interesantes es la generación de información de tendencias. Twitter extrae tendencias emergentes a partir de los tweets publicados y las mantiene a nivel local y nacional. Esto significa que a medida que una historia comienza a surgir, el algoritmo de temas de tendencia de Twitter identifica el tema en tiempo real. Este algoritmo en tiempo real se implementa en Storm como un análisis continuo de datos de Twitter "[53].

Pig

Pig es una plataforma de alto nivel para crear programas MapReduce [54] utilizados en Hadoop. El lenguaje de esta plataforma es llamado Pig Latin. Pig Latin abstrae la programación desde el lenguaje Java MapReduce en una notación que hace de MapReduce un lenguaje de programación de alto nivel, similar a la de SQL para sistemas RDBMS. Pig Latin puede ser ampliado utilizando UDF (Funciones Definidas por el Usuario) que el usuario puede escribir en Java, Python, Javascript, Ruby o Groovy y luego llamar directamente desde el lenguaje.

Pig fue desarrollado originalmente por Yahoo Research alrededor del 2006 por los investigadores para tener una forma ad-hoc de crear y ejecutar un trabajo map- reduce en conjuntos de datos muy grandes. En 2007, fue trasladado a *Apache Software Foundation.*

La capa de la infraestructura de Pig consiste en un compilador que produce programas de secuencias Map-Reduce, para los que ya existen implementaciones paralelas a gran escala (por ejemplo, la subproyectos Hadoop). Tiene las siguientes propiedades clave:

- La facilidad de programación. Es trivial para lograr la ejecución paralela de tareas y el análisis de datos simples. Las tareas complejas compuestas de múltiples transformaciones de datos relacionados entre sí están codificados explícitamente como secuencias de flujo de datos, lo que hace que sean fáciles de escribir, entender y mantener.
- Oportunidades de optimización. La forma en que se codifican las tareas permite que el sistema para optimizar su ejecución de forma automática, lo que permite al usuario centrarse en la semántica en lugar de la eficiencia.
- Extensibilidad. Los usuarios pueden crear sus propias funciones para realizar el procesamiento de propósito especial.

Yarn

Apache Hadoop YARN (por las siglas en inglés) es una tecnología de administración de clústeres [55]. YARN es una de las características clave de la segunda generación de la versión Hadoop 2 del marco de procesamiento distribuido de código abierto de Apache Software Foundation. Originalmente descrito por Apache como un gestor de recursos rediseñado, YARN se caracteriza ahora como un sistema operativo distribuido, a gran escala, para aplicaciones de big data.

En 2012, YARN se convirtió en un subproyecto del proyecto más grande Apache Hadoop. A veces llamado MapReduce 2.0, YARN es una reescritura de software que desacopla las capacidades de gestión de recursos y planificación de MapReduce del componente de procesamiento de datos, permitiendo a Hadoop soportar enfoques más variados de procesamiento, y una gama más amplia de aplicaciones. Por ejemplo, los clusters Hadoop ahora pueden ejecutar consultas interactivas y transmisiones de aplicaciones de datos de forma simultánea con los trabajos por lotes de MapReduce. La encarnación original de Hadoop empareja de cerca al sistema de archivos distribuidos Hadoop (HDFS) con el marco de programación MapReduce orientado a lotes, que se ocupa de la gestión de recursos y la planificación de tareas en los sistemas Hadoop, y soporta el análisis y la condensación de conjuntos de datos en paralelo.

YARN combina un administrador central de recursos que reconcilia la forma en que las aplicaciones utilizan los recursos del sistema de Hadoop con los agentes de administración de nodo que monitorean las operaciones de procesamiento de nodos individuales del clúster. Ejecutándose en clústeres de hardware básicos, Hadoop ha atraído un interés particular como zona de espera y de almacenamiento

de datos para grandes volúmenes de datos estructurados y no estructurados destinados al uso en aplicaciones de analítica. Separar HDFS de MapReduce con YARN hace al ambiente Hadoop más adecuado para las aplicaciones operativas que no pueden esperar para que terminen los trabajos por lotes.

Mahout

Apache Mahout es un proyecto de la Apache Software Foundation [56] para producir gratuitas implementaciones de distribución o de otra manera escalable de aprendizaje automático algoritmos se centraron principalmente en las áreas de filtrado colaborativo, agrupamiento y clasificación. Muchas de las implementaciones utilizan la plataforma Hadoop. Mahout también proporciona bibliotecas de Java para las operaciones matemáticas comunes (centrados en el álgebra lineal y estadísticas) y las colecciones de Java primitivas. Mahout es un trabajo en progreso; el número de algoritmos implementados ha crecido rápidamente.

A partir de la versión 0.10.0, el proyecto cambia su enfoque a la construcción de un entorno de programación backend-independiente, cuyo nombre en código "Samsara". El entorno se compone de un optimizador backend-independiente algebraica y un unificador algebraica DSL Scala en memoria y operadores algebraicos distribuidos. En el momento de escribir estas líneas plataformas soportadas son algebraicas Spark Apache y H2O, y Apache Flink. Soporte para MapReduce algoritmos se está eliminando gradualmente.

Mahout admite cuatro casos principales de uso de datos científicos:

- Filtrado colaborativo: Mira el comportamiento del usuario y hace recomendaciones de productos (por ejemplo, recomendaciones de Amazon).
- Agrupación: Toma los elementos de una clase en particular (como páginas web o artículos periodísticos) y los organiza en grupos naturales, de forma que los elementos pertenecientes al mismo grupo son similares entre sí.
- Clasificación: Aprende de las categorizaciones existentes y luego asigna los elementos no clasificados a la mejor categoría.
- Extracción de elementos frecuentes: Analiza elementos de un grupo (por ejemplo, elementos de un carrito de compras o términos en una sesión de consulta) e identifica qué elementos suelen aparecer juntos.

5

Big Data en la Hidroclimatología

5.1. ¿Qué es la hidroclimatología?

La hidroclimatología es un marco que nos permite analizar cómo la atmósfera ocasiona la variación espacio temporal de los elementos del ciclo hidrológico en escalas globales, regionales y locales. La hidroclimatología se definió cerca de 1967 como el estudio del clima sobre las aguas continentales. El ciclo hidrológico como un todo es el tema central de la hidroclimatología. En la actualidad la hidrología cumple un papel muy importante en el planteamiento del uso de los recursos hidráulicos, además tiene que ver con el suministro de agua, el drenaje de ríos y afluentes, así como con la protección contra la acción de los ríos.

5.1.1. Variables hidroclimatológicas

El análisis hidroclimatológico permite explicar cómo los cambios en las relaciones entre los elementos del sistema climático y los del ciclo hidrológico conducen a fenómenos cómo sequías, inundaciones y a cambios de largo plazo en la disponibilidad de recursos hídricos. Para determinar estos posibles cambios climáticos existen distintas variables hidroclimatológicas que nos ayudan a determinar el momento en que pueda ocurrir uno de estos cambios; es por ello que se va a dar una breve descripción de dichos parámetros a continuación.

5.1.1.1. Pluviosidad

Para explicar este parámetro, primero vamos a observar que es la lluvia.

Por lluvia se entiende el fenómeno atmosférico de tipo hidrometeorológico que se inicia con la condensación del vapor de agua contenido en las nubes [57].

Según la definición oficial de la Organización Meteorológica Mundial, la lluvia es la precipitación de partículas líquidas de agua, de diámetro mayor de 0,5 mm o de gotas menores, pero muy dispersas. Si no alcanza la superficie terrestre no sería lluvia, sino virga, y, si el diámetro es menor, sería llovizna. La lluvia se mide en milímetros. La lluvia depende de tres factores: la presión atmosférica, la temperatura y, especialmente, la humedad atmosférica. El agua puede volver a la tierra, además, en forma de nieve o de granizo. Dependiendo de la superficie contra la que choque, el sonido que producirá será diferente.

Ahora bien, la pluviosidad es la cantidad de lluvia que recibe un sitio en un período determinado de tiempo, donde el instrumento de medición se conoce como "pluviómetro". La medición se expresa en milímetros de agua y equivale al agua que se acumularía en una superficie horizontal e impermeable de 1 metro cuadrado durante el tiempo que dure la precipitación [58].

Un milímetro de agua de lluvia equivale a 1L de agua por m^2, que es otra forma de medir la cantidad de agua de lluvia. La clasificación de las precipitaciones [59] es la siguiente:

- **Débiles:** cuando su intensidad es $<= 2mm/h$
- **Moderadas:** $> 2mm/h$ y $<= 15mm/h$
- **Fuertes:** $> 15mm/h$ y $<= 30mm/h$
- **Muy fuertes:** $> 30mm/h$ y $<= 60mm/h$
- **Torrenciales:** $> 60mm/h$

5.1.1.2. Temperatura

Es un término utilizado para expresar la intensidad relativa del calor. Esta se identifica con la energía cinética de la translación de moléculas y de acuerdo con la teoría cinética de gases tiene su cero absoluto cuando la movilidad de las moléculas ha cesado totalmente. Esta también representa una habilidad de la materia de transferir calor a otros cuerpos. El calor no se transferirá si los cuerpos se encuentran a la misma temperatura. Existen tres escalas de medición de la temperatura: Fahrenheit, Centígrado (Celsius) y Kelvin.

5.1.1.3. Presión Atmosférica

Se conoce como presión atmosférica a aquella presión que ejerce el aire en cualquier punto de la atmósfera. El valor medio de la presión de la atmósfera terrestre es de 1013.25 hectopascales («hPa» es una unidad de presión que equivale a 100 pascales) o milibares a nivel del mar, la cual está medida a una latitud de 45°C. Entonces, cuando el aire está muy frío, lo que sucede con la atmósfera es que éste desciende y aumenta la presión lo cual lleva a presenciar un estado de estabilidad, dando lugar a lo que se llama anticiclón térmico y si por el contrario, el aire está muy caliente y asciende, baja la presión y provoca lo que se conoce como inestabilidad, formándose un ciclón o borrasca térmica.

Pero también puede pasar que esporádicamente suceda algo que no sucede con frecuencia que es que el aire caliente y el aire frío se mezclen, pero cuando ambos se encuentran en la superficie el aire frío empuja al aire caliente para arriba provocando que la presión descienda y aparezca un fenómeno de inestabilidad. La presión atmosférica también varía según la latitud. La menor presión atmosférica al nivel del mar se alcanza en las latitudes ecuatoriales.

5.1.1.4. Aforos

Es el método para medir un caudal. En Hidrología superficial puede ser necesario medir desde caudales (pocos litros/segundo) hasta grandes ríos con caudales de centenares o miles de $m3/seg$. Los aforos se dividen en:

- Aforos directos: Son los que se realiza con algún aparato o procedimiento para medir directamente el caudal.
- Aforos indirectos o continuos: se mide el nivel del caudal, y a partir de esta información con el nivel se estima el caudal. Este se realiza con el fin de saber el nivel diariamente o de modo continuo en diversos puntos de una cuenca, por eso también se conocen como aforos continuos.
- Aforos Químicos: Se emplea arrojando una sustancia química a un cauce, y este se diluye en la corriente, y aguas más abajo se toman muestras para analizarlas. Cuanto mayor es el caudal, mas diluida estará la muestra recogida.

El Ministerio de Agricultura y pesca, alimentación y medio ambiente de Madrid españa, tiene un sistema de información del Anuario de Aforos [60] que brinda los datos hidrológicos suministrados por las Confederaciones Hidrográficas y Administraciones Hidráulicas de Cuencas Intracomunitarias.

Uno de los objetivos del sistema es que los ciudadanos tengan libre acceso a la información. Dicho anuario proporciona los datos hidrológicos procedentes de las estaciones de aforo en ríos, embalses, conducciones y estaciones evapométricas asociadas a los embalses.

5.1.1.5. Evotranspiración

La evaporación es el proceso por el cual el agua pasa de fase liquida a fase de vapor, desde la superficie a la atmosfera. El agua puede evaporarse desde una gran variedad de superficies tales como suelos, lagos, ríos y vegetación húmeda. Este cambio de fase requiere un aporte de energía, proporcionado fundamentalmente por la radiación solar y en menor grado por el aire que circunda la superficie evaporante.

La transpiración consiste en la vaporización de agua líquida contenida en los tejidos de la planta y en el transporte del vapor de agua a la atmosfera. La transpiración depende, al igual que la evaporación, del suministro de energía para el cambio de fase, junto con el gradiente de presión circundante, que es la fuerza impulsora para el transporte de vapor a través de las estomas.

La evaporación y la transpiración ocurren simultáneamente y no hay forma sencilla de separar ambos procesos, por lo que al flujo de vapor de agua desde una cubierta vegetal se le denomina evapotranspiración. La proporción de evaporación y transpiración en un cultivo cambia según las diferentes fases de desarrollo predomina el suelo desnudo, y el principal proceso es el de evaporación.

La evapotranspiración se produce a través de la evaporación del agua presente en la superficie terrestre, junto con la que esta en mares, ríos y lagos y la que procede también de la tierra, incluyendo la transpiración de los seres vivos, en especial de las plantas. Como resultado de este proceso se determina la formación de vapor atmosférico, que, al llegar a las condiciones de condensación, retorna en parte a la superficie en forma de precipitación liquida o sólida.

Por tanto la evapotranspiración es la consideración conjunta de los procesos de evaporación y transpiración. La diferencia entre estos dos conceptos esta en la participación de los seres vivos en el segundo, que es el proceso físico a través del cual sus superficies pierden agua a la atmósfera mediante el proceso de transpiración.

5.1.1.6. Dirección del viento

Por lo general, la dirección del viento se define como la orientación del vector del viento en la horizontal. Para propósitos meteorológicos, la dirección del viento se define como la dirección desde la cual sopla el viento, y se mide en grados en la dirección de las agujas del reloj a partir del norte verdadero. Por ejemplo, un viento del oeste sopla del oeste, a 270° del norte. Un viento del norte sopla desde una dirección de 360°. La dirección del viento determina la del transporte de una pluma emitida.

El instrumento más común para medir la dirección del viento es **la paleta de viento**. Las paletas de viento señalan la dirección desde la cual este sopla. Pueden ser de formas y tamaños diferentes: algunas con dos platos juntos en sus aristas directas y dispersas en un ángulo (paletas separadas), otras con un solo platillo plano o una superficie aerodinámica vertical. Por lo general, son de acero inoxidable, aluminio o plástico. Al igual que con los anemómetros, se debe tener cuidado al seleccionar un sensor a fin de asegurar una durabilidad y sensibilidad adecuadas para una determinada aplicación.

La dirección del viento depende de la distribución y evolución de los centros isobáricos; se desplaza de los centros de alta presión (anticiclones) y su fuerza es tanto mayor cuanto mayor es el gradiente de presiones. La determinación de la dirección y velocidad del viento se realiza a partir del estudio de la distribución de la presión atmosférica en la geografía terrestre, es decir a partir de los mapas isobáricos, donde existen dos principios generales:

1. El viento va siempre desde los anticiclones a las borrascas.
2. Su velocidad se calcula en función de los juntas o separadas que estén las isobaras en el mapa. Cuanto más juntas estén las isobaras, más fuerza tendrá el viento y cuanto más separadas, menos.

5.1.1.7. Velocidad del viento

El viento produce energía porque está siempre en movimiento. Se estima que la energía contenida en los vientos es aproximadamente el 2 % del total de la energía solar que alcanza la tierra [61]. El contenido energético del viento depende de su velocidad. Cerca del suelo, la velocidad es baja, aumentando rápidamente con la altura. Cuanto más accidentada sea la superficie del terreno, más frenará ésta al viento. Es por ello que sopla con menos velocidad en las depresiones terrestres y más sobre las colinas. No obstante, el viento sopla con más fuerza sobre el mar que en la tierra.

Otras fuerzas que mueven el viento o lo afectan son la fuerza de gradiente de presión, el efecto Coriolis, las fuerzas de flotabilidad y de fricción y la configuración del relieve. Cuando entre dos masas de aire adyacentes existe una diferencia de densidad, el aire tiende a fluir desde las regiones de mayor presión a las de menor presión. En un planeta sometido a rotación, este flujo de aire se verá influenciado, acelerado, elevado o transformado por el efecto de Coriolis en cualquier parte de la superficie terrestre en la que nos encontremos. La creencia de que el efecto de Coriolis no actúa en el ecuador es un error: lo que sucede es que los vientos van disminuyendo de velocidad a medida que se acercan a la zona de convergencia intertropical y esa disminución de velocidad queda automáticamente compensada por una ganancia en altura del aire en toda la zona ecuatorial. A su vez, esa ganancia en altura da origen a la formación de nubes de gran desarrollo vertical y a lluvias intensas y prolongadas, ampliamente repartidas en la zona de convergencia intertropical, en especial en la zona ecuatorial. La fricción superficial con el suelo genera irregularidades en estos principios afectando al régimen de vientos.

La velocidad del viento se mide preferentemente en náutica en nudos y mediante la escala Beaufort: Esta es una escala numérica utilizada en meteorología que describe la velocidad del viento, asignándole números que van del 0 (calma) al 12 (huracán). Fue ideada por el Almirante Beaufort en el siglo XIX.

Escala	Denominación	Efectos observados	Nudos	Km/hora
0	Calma	El humo se eleva en vertical.	menos de 1	0 a 1,9
1	Ventolina ó brisa muy ligera	El viento inclina el humo, no mueve banderas.	1 a 3	1,9 a 7,3
2	Flojito ó brisa ligera	Se nota el viento en la cara.	4 a 6	7,4 a 12
3	Flojo ó pequeña brisa	El viento agita las hojas y extiende las banderas.	7 a 10	13 a 19
4	Bonancible ó brisa moderada	El viento levanta polvo y papeles.	11 a 16	20 a 30
5	Fresquito ó buena brisa	El viento forma olas en los lagos.	17 a 21	31 a 40
6	Fresco	El viento agita las ramas de los árboles, silban los cables, brama el viento.	22 a 27	41 a 51
7	Frescachón	El viento estorba la marcha de un peatón.	28 a 33	52 a 62
8	Duro	El viento arranca ramas pequeñas.	34 a 40	63 a 75
9	Muy duro	El viento arranca chimeneas y tejas.	41 a 47	76 a 88
10	Temporal ó tempestad	Grandes estragos.	48 a 55	89 a 103
11	Tempestad violenta	Devastaciones extensas.	56 a 63	104 a 118
12	Huracán	Huracán catastrófico.	64 y más	119 y más

Cuadro 5.1: Medición de la fuerza del viento según la escala Beaufort

5.1.1.8. Radiación solar

Es el flujo de energía que recibimos del Sol en forma de ondas electromagnéticas de diferentes frecuencias (luz visible, infrarroja y ultravioleta). Aproximadamente la mitad de las que recibimos, comprendidas entre 0.4m y 0.7m, pueden ser detectadas por el ojo humano, constituyendo lo que conocemos como luz visible. De la otra mitad, la mayoría se sitúa en la parte infrarroja del espectro y una pequeña parte en la ultravioleta. La porción de esta radiación que no es absorbida por la atmósfera, es la que produce quemaduras en la piel a la gente que se expone muchas horas al sol sin protección. La radiación solar se mide normalmente con un instrumento denominado piranómetro [62].

En función de cómo reciben la radiación solar los objetos situados en la superficie terrestre, se pueden distinguir estos tipos de radiación:

- Radiación directa: Es aquella que llega directamente del Sol sin haber sufrido cambio alguno en su dirección. Este tipo de radiación se caracteriza por proyectar una sombra definida de los objetos opacos que la interceptan.
- Radiación difusa: Parte de la radiación que atraviesa la atmósfera es reflejada por las nubes o absorbida por éstas. Esta radiación, que se denomina difusa, va en todas direcciones, como consecuencia de las reflexiones y absorciones, no sólo de las nubes sino de las partículas de polvo atmosférico, montañas, árboles, edificios, el propio suelo, etc. Este tipo de radiación se caracteriza por no producir sombra alguna respecto a los objetos opacos interpuestos. Las superficies horizontales son las que más radiación difusa reciben, ya que ven toda la bóveda celeste, mientras que las verticales reciben menos porque sólo ven la mitad.
- Radiación reflejada: La radiación reflejada es, como su nombre indica, aquella reflejada por la superficie terrestre. La cantidad de radiación depende del coeficiente de reflexión de la superficie, también llamado albedo. Las superficies horizontales no reciben ninguna radiación reflejada, porque no ven ninguna superficie terrestre y las superficies verticales son las que más radiación reflejada reciben.
- Radiación global: Es la radiación total. Es la suma de las tres radiaciones. En un día despejado, con cielo limpio, la radiación directa es preponderante sobre la radiación difusa. Por el contrario, en un día nublado no existe radiación directa y la totalidad de la radiación que incide es difusa. Los distintos tipos de colectores solares aprovechan de forma distinta la radiación solar. Los colectores solares planos, por ejemplo, captan

la radiación total (directa + difusa), sin embargo, los colectores de concentración sólo captan la radiación directa. Por esta razón, los colectores de concentración suelen situarse en zonas de muy poca nubosidad y con pocas brumas, en el interior, alejadas de las costas.

5.1.1.9. Presión Barométrica

El peso del aire que forma nuestra atmósfera ejerce una presión sobre la superficie de la tierra. Esta presión es conocida como presión atmosférica. Generalmente, cuanto más aire hay sobre una zona más alta es la presión atmosférica, esto significa que la presión atmosférica cambia con la altitud. Por ejemplo, la presión atmosférica es mayor a nivel del mar que en la cima de una montaña.

Para compensar esta diferencia y facilitar la comparación entre lugares con distintas altitudes, la presión atmosférica generalmente se ajusta a la presión equivalente a nivel del mar. Esta presión es conocida como presión barométrica. En realidad la Vantage Pro2 mide la presión atmosférica. Cuando usted introduce la altitud de su localización en la modalidad Setup (Configuración), la Vantage Pro2 guarda el valor de compensación necesario para traducir constantemente la presión atmosférica a presión barométrica.

La presión barométrica también cambia con las condiciones climáticas locales, lo que la convierte en una herramienta para pronosticar el tiempo extremadamente importante y útil. Las zonas de alta presión generalmente son asociadas con buen tiempo, mientras que las de baja presión son asociadas con mal tiempo. Sin embargo, para fines de pronóstico, generalmente el valor absoluto de presión barométrica es menos importante que el cambio en la misma. En general, un aumento de presión indica un mejoramiento de las condiciones climáticas, mientras que una caída de presión indica deterioro de las mismas [63].

5.1.2. Ríos

El término río es una palabra que procede de la voz del latín **rius**. Por definición, un río es una corriente natural formada por agua dulce que fluye continuamente. Puede desembocar o morir en un lago, en el mar o en otro río [64]. En este último caso recibe el nombre de afluente y el punto de unión de ambos se llama confluencia.

5.1.2.1. Partes de un río

Desde su nacimiento hasta la desembocadura, un río pasa por distintas etapas o partes diferentes. Cada río, en función de su naturaleza y geografía es distinto, pero normalmente suelen tener en común las siguientes partes:

- **Curso alto:**

  El curso alto de un río o de gravedad alta es aquella parte más montañosa o escarpada. Es la zona donde las pendientes suelen ser más pronunciadas e inclinadas. Aquí se encuentra el nacimiento y la cabecera del río. En esta parte del río el agua suela bajar con cierta velocidad, con alta capacidad de erosión del terreno, y puede arrastrar pequeñas piedras y rocas. Por tanto, al principio, donde el terreno tiene mucha pendiente, el río corre velozmente arrancando del fondo y de los lados tierras y piedras.

  En esta área de algunos ríos se pueden formar los rápidos. Que es donde el agua circula por una pendiente algo mayor de lo habitual, aumentando su turbulencia y velocidad de forma considerable. También se pueden encontrar saltos de agua, cataratas o cascadas.

- **Curso medio:**

  El curso medio de un río o de gravedad inestable es la zona de llanura por la que discurre. Las aguas bajan más calmadas y con una velocidad menor que en el curso alto. En esta parte del río, se arrastran los materiales que han sido erosionados. Aquí se pueden formar meandros y encurvamientos para esquivar o rodear los grandes obstáculos que encuentra a su paso.

  Esta zona también es dónde se le pueden unir otros ríos, que cómo ya hemos dicho antes, se les llama afluentes. Aunque también puede haber en el curso alto, pero son más pequeños. Normalmente, al principio del curso medio de un río se suelen construir embalses, presas o centrales hidroeléctricas. Aunque esto depende mucho de cada caso en particular.

- **Curso bajo:**

  El curso bajo de un río es la parte final, cuando desemboca o muere en el mar. En este punto el cauce del río se ensancha y el agua fluye a poca velocidad. En esta zona, al circular el agua dulce muy lentamente, se van sedimentando o se depositan todos los materiales que ha ido arrastrando desde el curso alto.

5.1.2.2. Tipos de ríos

Existen diferentes tipos de ríos y se pueden clasificar en base a su actividad, caudal, geometría, morfología o composición de las aguas. Pero también por factores como la cantidad de curvas o meandros, divisiones o bifurcaciones con las que cuenta. Vamos a verlos:

- **Ríos estacionales:** Son los que están ubicados en zonas donde las estaciones son muy diferentes entre sí. Alternando temporadas de sequía y de lluvias. Por tanto presentan grandes diferencias de caudal en función de la estacionalidad. Suelen encontrarse en zonas de alta montaña, pero también en zonas bajas, aunque son menos habituales.
- **Ríos perennes:** Son los que se suelen ubican en zonas con grandes precipitaciones. No es habitual que presenten grandes cambios de caudal durante el año, ya que cuentan con un aporte de agua constante. Los ríos perennes también surgen de corrientes subterráneas, por lo que no siempre es necesario que se encuentre en una región de precipitaciones regulares.
- **Ríos Alócatenos:** Son aquellos ríos que atraviesan zonas muy secas, áridas o incluso desérticas. Esto es así ya que su nacimiento se encuentra a muchos kilómetros y es una zona muy lluviosa o húmeda. Dos buenos ejemplos son el río Colorado en USA o el Nilo.
- **Ríos transitorios:** Se encuentran en zonas de clima desértico o muy seco. Su caudal es tremendamente variable. Puede fluir libremente durante varios kilómetros y posteriormente desaparecer durante varios meses, volviendo a surgir cuando caen fuertes lluvias. Este tipo de río representan un gran peligro, ya que cuando llueve fuertemente, pueden reaparecen con gran violencia en forma de fuertes riadas.
- **Ríos meandriformes:** Por norma general, tiene un canal único pero que en su discurrir, tiene o forma un gran número de meandros. Adoptando así una forma muy sinuosa que lo caracteriza. Los ríos meandriformes cuentan con una característica especial. Y es que debido a que sus aguas tienen una velocidad variable por las curvas, son capaces de erosionar el terreno y a la vez crear zonas de sedimentos.
- **Ríos anatomosados:** Se caracterizan por contar con diferentes canales y son capaces de transportar gran cantidad de sedimentos y materiales. Al tener poca energía, si se topan con un obstáculo, en lugar de erosionarlo, los rodean si es posible. Los ríos anatomosados pueden formar en ciertas ocasiones islas sedimentarias.

- **Ríos rectilíneos:** Este tipo de río están formados por un canal principal y algunas pequeñas bifurcaciones inestables. Normalmente y debido a su naturaleza, estas corrientes de agua dulce suelen tener una gran potencia y capacidad de erosión.

- **Embalses**: Un embalse es un lago artificial o un gran cuerpo de agua dulce. Mucha gente piensa en un depósito como un lago e incluso podrían usar las palabras de manera intercambiable. Sin embargo, la diferencia clave es que los embalses son artificiales y hecho por los seres humanos, mientras que los lagos son cuerpos de agua de origen natural. Los embalses son grandes a la fuerza, ya que tienen que almacenar el suficiente agua para abastecer a núcleos urbanos.

5.1.2.3. Tipos de embalses

Hay dos tipos principales de los embalses:

- **Embalses de valle represado:** se crean en los valles entre montañas. Por lo general, hay un lago existente o cuerpo de agua en el lugar de la construcción. Las laderas de las montañas se utilizan como las paredes del depósito para contener el agua. Se construye un dique o muro artificial en el depósito en el punto más estrecho para mantener el agua.

  Para que el embalse de valle represado pueda construirse, el río que llenará el depósito debe desviarse, por lo que el suelo se puede borrar para sentar una base para la presa. A continuación se pone un revestimiento de hormigón y se puede empezar con la construcción de la presa. Pueden pasar años hasta que se construya la presa, pero una vez construida representa una fuente constante de agua.

  En ocasiones también se puede desviar el agua de ríos o arroyos locales a un embalse existente. Aunque esto se puede aplicar a muchas áreas geográficas diferentes, resulta más problemático que un embalse de valle represado. El río Támesis en Londres es un ejemplo de este tipo de embalses.

- **Depósitos de aguas:** aquí el agua se almacena en tanques de de hormigón por encima o por debajo del suelo. Este tipo de depósitos deben estar en un lugar de mayor altura para permitir que el agua fluya a donde tiene que ir. Claramente, los embalses son un gran recurso para el almacenamiento de agua necesaria cuando otras fuentes son escasas. Sin embargo, los depósitos también pueden servir para otros fines. Dado que el depósito de agua está relativamente quieto, sino que también se puede utilizar para ayudar a limpiar el agua antes de desembocar en una planta de tratamiento de agua para el consumo humano.

5.1.3. Definición de Cuenca Hídrica

Se entiende por cuenca a aquella depresión o forma geográfica [65] que hace que el territorio vaya perdiendo altura a medida que se acerca al nivel del mar. Las cuencas hidrográficas son aquellas que hacen que el agua que proviene de las montañas o del deshielo, descienda por la depresión hasta llegar al mar. En algunos casos, la cuenca puede no alcanzar el nivel del mar si se trata de un valle encerrado por montañas, en cuyo caso la formación acuífera será una laguna o lago.

Las cuencas hidrográficas pueden ser divididas en dos tipos principales: las cuencas endorreicas, aquellas que no llegan al mar, que tienen como resultado la formación de sistemas de agua estancada (como lagos o lagunas); y las cuencas exorreicas, aquellas que sí llegan al mar y que por lo tanto no quedan encerradas entre los diferentes conjuntos de montañas. Normalmente, las cuencas, tanto sean endorreicas o exorreicas pueden generar un gran número de afluentes que caen todos en el curso de agua principal, ya sea mar, océano, lago o laguna. Al mismo tiempo, a medida que esos afluentes se acercan a su destino final van perdiendo la intensidad original que tenían al comenzar su curso de descenso.
Las cuencas hidrográficas son de gran importancia para el medio ambiente así como también para el ser humano. En este sentido, actúan como importantes reservorios de agua que pueden ser aprovechadas no sólo por el ser humano para su consumo personal, diferentes actividades económicas como la agricultura o la navegación, sino también para el consumo de los animales y plantas y por tanto el desarrollo de sistemas bióticos completos y duraderos.

De más está decir que en el planeta Tierra encontramos numerosas cuencas hidrográficas, teniendo cada una de ellas características particulares. Algunos de los mares actuales se consideran cuencas hidrográficas endorreicas debido a la progresiva pérdida de su contacto con el océano.

Subcuencas: Es toda área que desarrolla su drenaje directamente al curso principal de la cuenca. Varias subcuencas pueden conformar una cuenca.

Canales: Un canal es un estrecho curso de agua, de origen natural o artificial. Son relativamente fáciles de reconocer ya que no tienen la amplitud ni el gran volumen de agua de los lagos o de los ríos, y sus aguas no son tan rápidas. Los canales artificiales son muy comunes en muchas ciudades; quizá la más famosa de ellas es Venecia, cuyos canales funcionan como calles por donde circulan múltiples embarcaciones de transporte.

Existen canales naturales y canales artificiales. Los primeros son aquellos accidentes geográficos efectuados por la naturaleza sin la intervención del hombre y localizados en los últimos tramos de un río, un delta o un estrecho, aunque este ya suele ser bastante angosto. Los canales artificiales son también pasajes estrechos, pero pasan a través de una divisoria de aguas, la región límite entre dos cuencas hidrográficas. Para tener un canal artificial es necesario cavar una larga zanja y asegurar su suministro continuo de agua; esto suele lograrse conectándolo directamente con el mar, tomando agua de ríos o manantiales o bombeando el líquido de otras fuentes. Por lo general, los canales sirven para conectar cuerpos de agua como lagos, ríos, mares u océanos.

Las razones para construir canales son variadas, pero por lo general sirven para conectar cuerpos de agua como lagos, ríos, mares u océanos. Por ejemplo, el Canal de Suez fue creado en el siglo XIX para separar físicamente Asia y África y así facilitar el paso desde Europa hasta el sur de Asia, conectando el mar Mediterráneo con el mar Rojo.

Tipos de canales artificiales

Los canales artificiales pueden ser de 2 tipos: vías navegables y acueductos. Las vías navegables se utilizan primordialmente para transportar personas y mercancía a bordo de las embarcaciones, ya sea dentro de las ciudades o como conexión entre cuerpos de agua naturales (ríos, lagos, etcétera). Los acueductos conducen el agua hacia lugares específicos, generalmente poblaciones humanas que requieren el agua para su consumo.

Por otra parte, los canales de energía (Power canal) son construidos con el propósito de generar con ellos energía hidráulica, con la que asimismo puede generarse energía eléctrica. Fueron muy comunes en Nueva Inglaterra, Estados Unidos, durante la Revolución Industrial por su capacidad para proveer energía a los edificios y fábricas de molinos.

5.2. Análisis de Datos Hidroclimatológicos

Los volúmenes de datos climatológicos presentan un rápido crecimiento debido al desarrollo de las nuevas plataformas de adquisición y a los avances de las tecnologías de almacenamiento. Tales avances proveen nuevos retos para los métodos de análisis de datos y como resultado hay un interés creciente en la aplicación y desarrollo de nuevos métodos de Machine Learning y minería de datos en el análisis de datos climatológicos [66].

Por otra parte y debido a la alta interacción del agua con los componentes atmosféricos de la tierra se hace necesario administrarla (en sus diferentes fases) en un modelo sostenible a través de la gestión integrada de los recursos hídricos. El agua es más que nunca un elemento importante en la formación y conservación del medio ambiente y por ello está recibiendo una atención global. El ambiente de la tierra se compone de varios sistemas que trabajan juntos y en armonía pero con algunas relaciones e interacciones complejas. Para un enfoque sistémico de la hidrología y la hidroclimatología es necesario tener un claro entendimiento de los diferentes componentes del sistema de la tierra y del posicionamiento de cada subsistema con respecto a los demás. Esto ayudará a comprender y proveer una interpretación física de los diferentes eventos climáticos es escalas regionales y globales.

El enfoque tradicional de la hidrología y la hidroclimatología se enfoca sobre esas variables que son más tangibles y representativas como lo son la temperatura y sus variaciones, el viento cerca de la superficie de la tierra, la humedad, las características de las nubes, la precipitación en sus diversas formas y la varidedad estacional de la escorrentía. Sin embargo esas variables podrian no definir completamente los sistemas hidrológicos e hidroclimatológicos y sus variaciones que dependen de diferentes parámetros tales como la estructura vertical de la atmósfera, la subyacente influencia de la tierra y el mar, las actividades antropogénicas y muchos otros factores que no son totalmente explorados en los cursos clásicos de hidrología [67]. El ciclo hidrológico es un modelo de naturaleza holística y ha sido citado como un sistema verdadero en la literatura.

La captura de las variables mencionadas se realiza por medio de redes y estaciones de monitoreo hidroclimatológicos que cuentan con instrumentos y sensores sofisticados que permiten hacer mediciones con mayor precisión, donde la transmisión de los datos se lleva a cabo en tiempo real lo cual habilita y facilita la toma de decisiones a través del suministro de información oportuna, abundante y confiable.

Las redes de monitoreo se componen de un conjunto de estaciones metereológicas e hidrometeorológicas dispuestas para enviar información a una estación central. Las estaciones meteorológicas son sistemas de monitoreo con almacenamiento interno de datos que pueden ser configuradas para transmitir los datos vía radiofrecuencia y/o sistema celular GPRS. Los datos que se envían están relacionados con la precipitación, la temperatura, humedad relativa, radiación solar, velocidad del viento, dirección del viento, presión barométrica y evotranspiración. De otro lado se encuentran las estaciones hidrometeorológicas que están diseñadas para realizar mediciones periódicas (cada 2, 5, 10, 30 minutos o lo que se requiera) de los niveles de agua en ríos y quebradas, además de medir las variables meteorológicas. El sistema central permite la recolección, almacenamiento y procesamiento de información que se puede vincular a aplicaciones web para consulta remota desde cualquier parte del planeta. A continuación se relaciona la figura No. 5.2 que muestra el esquema de una red hidroclimatológica.

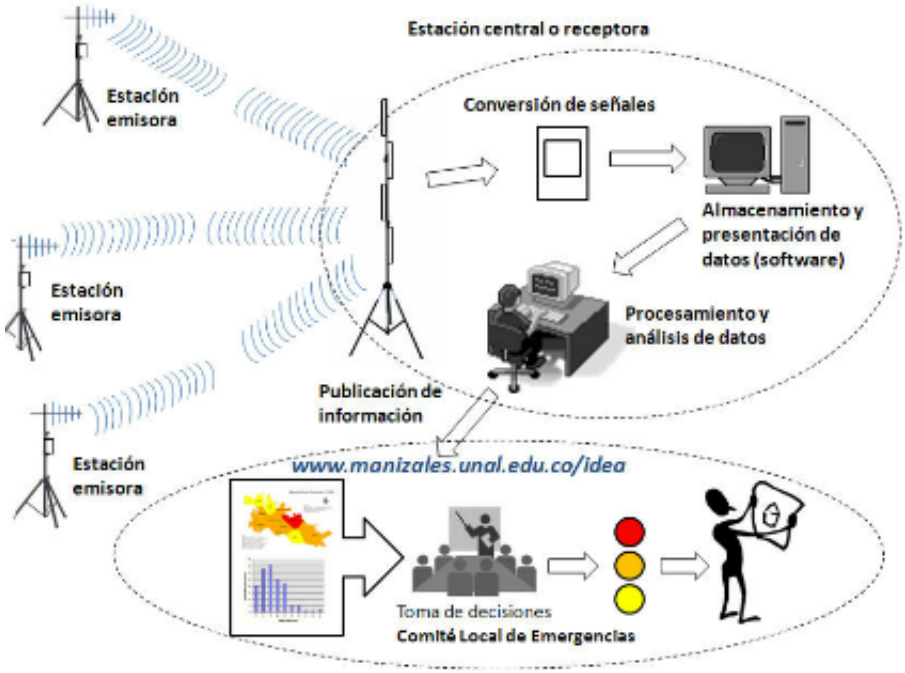

Figura 5.1: Estructura de la red de monitoreo, fuente [68]

5.2.1. Descripción de Algunas Técnicas Utilizadas

En general y teniendo presente que el analisis de datos hidroclimatológicos se basa principalmente en la matemática y la estadística, dadas las características de los datos hidroclimatológicos, todas las técnicas de análisis descritas en el capítulo anterior (e incluso las que no se describieron) son aplicables al análisis del tipo de datos en cuestión, dado que estos cuentan con los atributos Big Data como volumen, variedad, velocidad, veracidad, valor, variabilidad, visibilidad, veredicto y validez.

A la hora de elegir cuál técnica usar, se debe establecer cuál es el cuestionamiento que se tiene y qué tipo de respuesta se está buscando, por ejemplo, en [69] para determinar el balance hidrológico en cuencas hidrográficas, utilizaron el modelo de las llamadas "ecuaciones diferenciales" de Lvovitch cuya ecuación se expresa así:

$$P = Q + E_A \pm \Delta S \tag{5.1}$$

donde: P = precipitación, Q = escurrentía, E_A = evapotranspiración real, ΔS = cambios en el almacenamiento de agua.

Por otro lado, en [70, Pág. 43] proponen 4 pasos principales para la clasificación de la lluvia extrema. En primer lugar, se extrajo un conjunto de 17 variables que describen una gama de características de precipitación extrema para cada serie de tiempo. Segundo, el análisis por componentes principales (ACP, también conocido como el análisis empírico de la función ortogonal) fue usado para reducir la dimensionalidad de las variables agrupadas, identificando los componentes más importantes de la variabilidad dentro del conjunto de datos. En tercer lugar, las regiones de precipitaciones extremas se clasificaron a partir del análisis de agrupaciones en las puntuaciones de PCA y ubicación geográfica (usando k-means). Por último, se produjeron los mapas de las regiones clasificadas.

5.2.1.1. Decimated Wavelet Transform

La transformada de onduleta decimada discreta es ampliamente aplicada en predicciones de series de tiempo basadas en ondas y está basada en el algoritmo piramidal. La transformada se realiza en dos partes: el componente de descomposición que transforma las series de tiempo originales en conjuntos de aproximación y coeficientes detallados, cada cual con niveles detallados de resolución; y el componente de reconstrucción que sirve para recrear las series de tiempo originales a partir de dichos conjuntos de coeficientes. Esta técnica es muy útil para la supresión de señales y eliminación de ruido en los datos, dado que ayuda a eliminar componentes que no aporten información necesaria en la

señal, de igual manera se pueden recortar las amplitudes de las señales de alta frecuencia que representen ruidos o perturbaciones indeseables. Por otra parte, la transformada Wavelet permite realizar la comprensión de señales, la detección de autosimilitudes (estructuras que se repiten por doquier y a cualquier escala de la señal), la detección de discontinuidades y puntos de falla.

Como se anota en [66, Pág. 13] se debe prestar atención a la hora de implementar esta técnica de análisis ya que una aplicación incorrecta de la transformada de onduleta decimada discreta, ocacionará predicciones erróneas.

5.2.1.2. Modelo de Caracterización de Cuencas

El modelo de caracterización de cuencas (Basin Characterization Model) es un modelo sencillo basado en la red que calcula el balance hídrico para cualquier paso temporal o escala espacial utilizando insumos climáticos, precipitación y temperatura mínima y máxima del aire. La evotranspiración potencial se calcula a partir de la radicación solar con sombreado topográfico y nubosidad, la nieve se acumula, se sublima y se derrite, y el exceso de agua se mueve a través del contorno del suelo, cambiando el almacenamiento del agua en el mismo. Los cambios en el agua del suelo se utilizan para calcular la evotranspiración real y cuando se resta de la evotranspiración potencial se calcula el déficit hídrico. Dependiendo de las propiedades del suelo y la permeabilidad del lecho rocoso subyacente, el agua puede convertirse en recarga o escurrimiento. El enrutamiento se realiza a través del post- procesamiento para estimar el flujo base, el flujo de agua y la recarga de agua subterránea. Los insumos climáticos escalados mensualmente y las variables de salida hidrológicas se pueden examinar por cualquier polígono de un tamaño que represente regiones o cuencas hídricas, o la distribución a través del paisaje [71].

Cabe anotar que este modelo fue usado para calcular un balance de agua para la región hidrológica de California en los Estados Unidos, que incluye todas las cuencas que drenan el estado. El esquema anteriormente descrito se muestra en la figura No. 5.2.

5.2.2. Caso de Estudio

Emilcy Juliana Hernández Leal en [72] propone un modelo genérico para la administración de Big Data en el dominio de datos ambientales, donde hizo una identificación de cada una de las capas que hacen parte de la propuesta así como nombra las tecnologías a utilizar en cada una de ellas para finalmente definir el modelo específico para los datos ambientales.

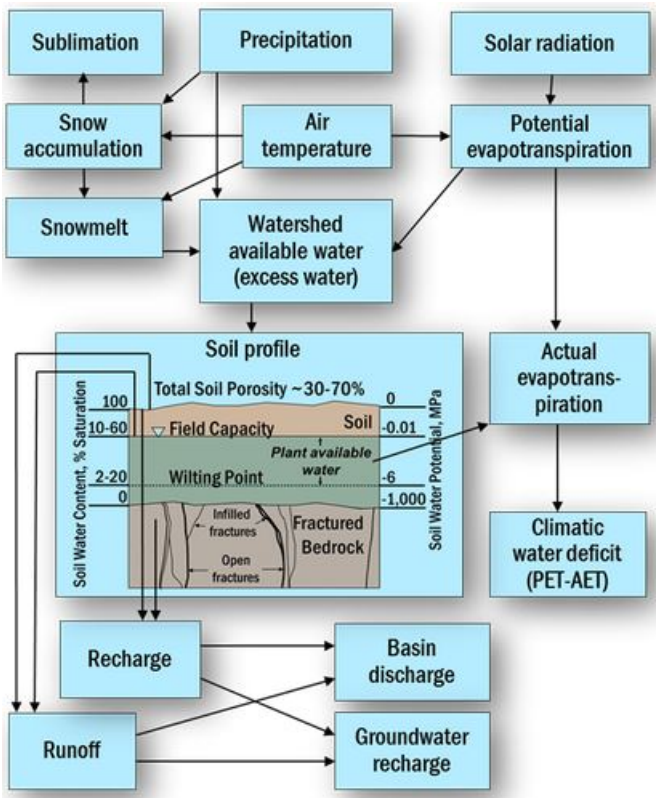

Figura 5.2: Relaciones entre los componentes del modelo de caracterización de cuencas, fuente [71]

Referente a la identificación de la capas asociadas al modelo de administración de Big Data, se debe:

1. Hacer el descubrimiento de los datos, definir los datos de interés, encontrar las fuentes desde donde se tomarán (históricos en textos planos, generados en tiempo real, almacenados en bases de datos, obtenidos a partir de sensores, entre otros), llevar estos datos a un esquema que pueda interactuar con el sistema y determinar cómo serán tratados.

2. Hacer un proceso de extracción y limpieza de los datos.

3. Estructurar los datos para su posterior análisis, esto incluye la creación de una estructura lógica para los conjuntos de datos tratados, almacenar estos datos bajo el medio elegido y empezar algunos análisis para hallar relaciones, alternativas, patrones, etc.

4. Realizar el procesamiento de los datos por medio de la aplicación de algoritmos, de procesos estadísticos, técnicas de minería, entre otras.

5. Interpretación de los datos obtenidos.

El esquema propuesto se muestra en la figura No. 5.3.

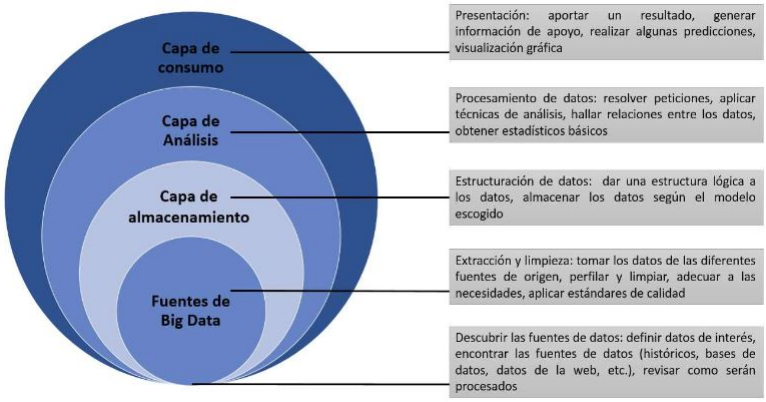

Figura 5.3: Capas del modelo genérico para la administración de Big Data, fuente [72]

5.2.2.1. Fuentes de Datos Ambientales

En esta capa, la autora considera como fuente del tipo de datos ambientales, los resultantes de monitoreo realizado a través de las redes hidroclomatológicas y además añade: "El crecimiento en el número de datos hidrometeorológicos generados por una red de monitoreo es considerable; por ejemplo, si se piensa en una estación que genera un registro por variable cada 5 minutos, en un día son aproximadamente 288 registros por variable por estación; al pasar a una red de 20 estaciones se lograrían 5.760 registros en un día por variable en la red; si cada estación de la red mide 7 variables se tendría 40.320 registros al día en la red. Pero si en la ciudad hay 4 redes cada una con las 20 estaciones, se estaría hablando de 80 estaciones que generarían al día 161.280 datos, en una semana se ascendería a 1'.128.960 y en un año, la cifra alcanzaría los 58'.705.920 datos; y si se va más lejos, en una década se obtendrían alrededor de 587 millones de datos" [72, Pág. 82].

5.2.2.2. Proceso de ETL

Para el proceso de extracción, transformación y carga se deben tener en cuenta las actividades en la tarea de filtrado: el filtrado de detección y el filtrado de fallas. El primero se enfoca en la detección de errores presentes en las mediciones de cada una de las variables. El segundo, recibe los datos con los errores detectados y los organiza siguiendo un estándar definido.

5.2.2.3. Almacenamiento

En esta capa, Emilcy propone el uso de bases de datos híbridas dada la oportunidad que se tiene en la actualidad de combinar las tecnologías de almacenamiento tradicional con las tecnologías de *datawarehouse* (Bodegas de Datos que permiten mayor escalabilidad), así como el uso de bases de datos no relacionales (NoSQL).

5.2.2.4. Análisis de Datos

De acuerdo a la versatilidad que brinda Big Data en esta capa, existe la posibilidad de aplicar consultas relacionales, minería de datos, machine learning, Deep learning, análisis estadístico, análisis dimensional, técnicas de clasificación, técnicas de regresión, análisis predictivo y otras técnicas de inteligencia artificial, lo cual es acorde con lo que hemos descrito en esta investigación. En la capa de análisis de datos se debe tener un claro entendimiento del conocimiento que se desea adquirir para la correcta elección de/las técnicas a implementar sobre los conjuntos de datos.

5.2.2.5. Presentación de Datos

Se deben presentar al publico objetivo tanto los datos históricos como los datos capturados en tiempo real, dado que los datos hidrometeorológicos históricos son fundamentales para entender la variabilidad climática que caracteriza una región y periodo de tiempo.

Implementación del modelo propuesto

En esta fase, la autora empleó una técnica de extracción, transformación y carga de datos para garantizar la integridad de los mismos. Aplicó la técnica de análisis *deep learning* para la predicción del comportamiento de las variables precipitación, temperatura, humedad y presión; además implementó algoritmos

de clustering (K-means y Canopy) utilizando **Mahout** en **Hadoop** [72, Pág. 87]. En el análisis de los datos, se utilizó una red neuronal perceptrón multicapa y las pruebas fueron efectuadas sobre un clúster para procesamiento paralelo masivo compuesto por 72 nodos.

Los datos hidrometeorológicos fueron obtenidos por el Instituto de Estudios Ambientales (IDEA) de la Universidad Nacional de Colombia, sede Manizales.

A continuación se relacionan las imágenes con los resultados de la implementación del los algoritmos k-means y Canopy.

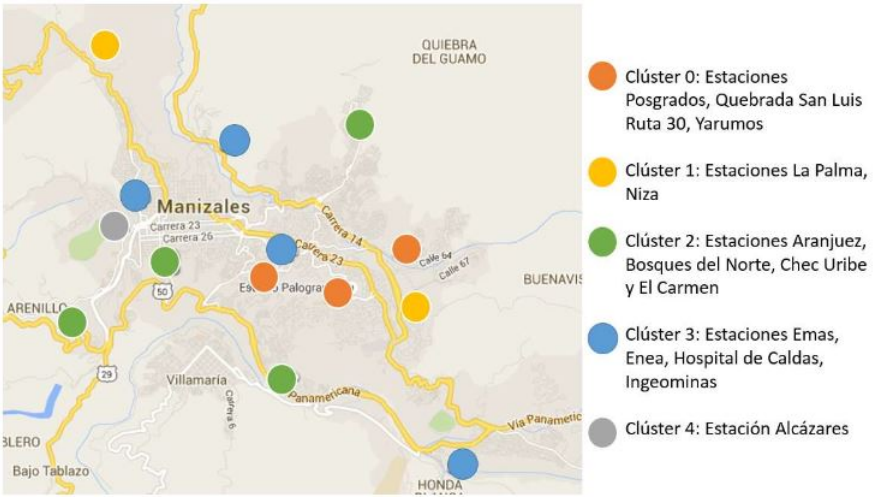

Figura 5.4: K-means con K=5, fuente [72, Pág. 104]

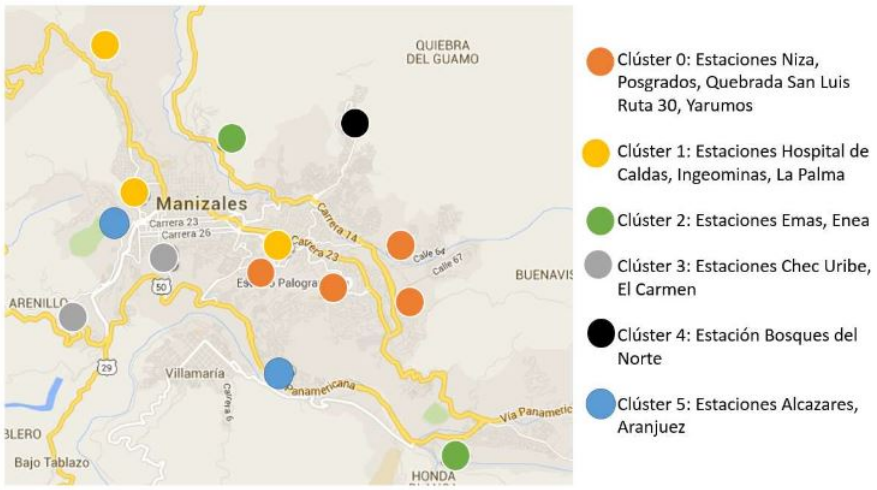

Figura 5.5: Resultado de Clustering con Canopy, fuente [72, Pág. 106]

En cuanto a la forma de presentar los resultados del análisis, Emilcy utilizó la presentación de datos en tiempo real a través de la página del Instituto (idea.manizales.unal.edu.co), como se muestra en las figuras 5.6 y 5.7.

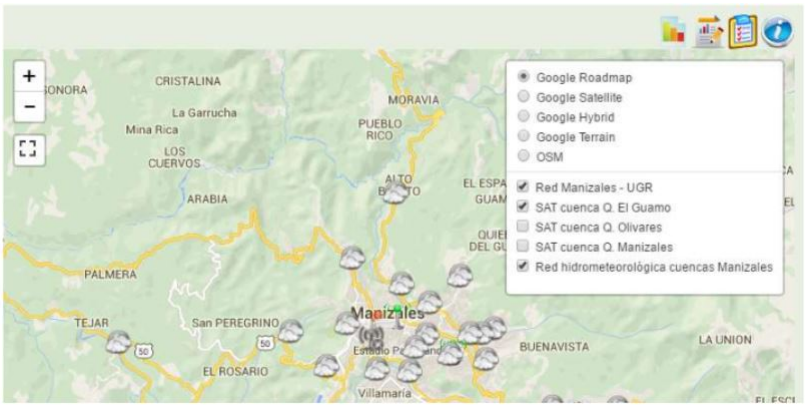

Figura 5.6: Visualización de las capas en el aplicativo de estado del tiempo., fuente [72, Pág. 107]

Figura 5.7: Presentación de datos detallados de cada estación en tiempo real, fuente [72, Pág. 108]

5.2.3. Visualización de Datos y su Importancia en la Hidroclimatología

En este punto surge un interrogante: ¿por qué es importante la visualización de datos?. Porque de la forma en la que la mente humana procesa la información usando gráficos para visualizar ingentes cantidades de datos complejos, es más fácil que examinarlos a través de hojas de cálculo o informes. La visualización de datos es la forma más fácil y rápida de transmitir conceptos de manera universal y se puede experimentar con diferentes escenarios haciendo pequeños ajustes [73]; la visualización de datos promueve la exploración creativa de éstos.

Ahora bien, teniendo en cuenta que son tan variados los tipos de datos hidroclimatológicos que se captan por medio de las estaciones de telemetría y los sensores, poder presentarlos de forma clara es un reto que las técnicas de visualización ayudan a abordar.

5.2.3.1. Framework de Visualización: Principios y Diseño

Los sistemas de **visualización** basados en computadores proveen una representación de conjuntos de datos diseñados para ayudar a las personas a realizar las tareas de una manera más efectiva.

La visualización es adecuada cuando existe la necesidad de aumentar las capacidades humanas en vez de reemplazar a las personas por métodos de toma de decisiones computacionales. Los diseñadores de visualizaciones deben considerar tres tipos de limitaciones: la capacidad computacional, la capacidad cognitiva y de percepción de las personas y la capacidad de las pantallas. La utilización de la visualización puede ser analizado en términos de **por qué** los usuarios la necesitan, **qué** datos serán mostrados y **cómo** la codificación visual y la interacción de los modismos son construidos en términos de opciones de diseño [74].

Este diseño y análisis proporciona un framework sistemático y compresivo (diseñado por Tamara Munzner, científica Informática - Stanford) sobre la visualización en términos de principios y opciones de diseño que abarca técnicas de visualización científica de información para datos abstractos, espaciales y técnicas visuales para entretejer la transformación y el análisis de datos con la exploración visual interactiva. Este framework hace énfasis de la validación cuidadosa de la *eficacia* y la consideración de la *función* antes que la forma.

5.2.3.2. Tecnologías para la visualización de datos

En la actualidad existen un sin numero de herramientas, pero deseamos nombrar una que está teniendo un impacto positivo en el análisis de datos hidroclimatológicos por la capacidad de abstracción que ofrece y la fácil usabilidad que proporciona: Tableau.

Tableau es un software para la visualización de datos de gran utilidad que trabaja sobre el enfoque de arrastrar y soltar campos de una tabla para realizar análisis visual de esos datos.

De otro lado, se tienen lenguajes como **R**, que es un entorno de programación para el análisis estadístico y gráfico, **D3.js** que es una librería de JavaScript para producir, a partir de datos, inforgramas dinámicos e interactivos en los navegadores web y **Python** que actualemente cuenta con librerías para la visualización, como por ejemplo, matplotlib, geoplotlib, pandas, etc.

5.3. Beneficios del análisis Big Data aplicado a la hidroclimatología

Las tecnologías con las que se cuenta para Big Data son de gran ayuda en el procesamiento de los datos, dado que se trabaja sobre lenguajes de programación de alto nivel, que operan sobre sistemas de archivos distribuidos, lo cual mejora enormemente la capacidad de análisis en tiempo real, puesto que permite a las aplicaciones trabajar con miles de nodos (de bajo costo) y petabytes de datos, sobre la filosofía de "llevar la computación a los datos y no inversamente". Lo que se logra con ello, es minimizar costos en el análisis de la gran cantidad de datos hidroclimatológicos.

Además, con la adopción de la computación en la nube, para las organizaciones es mucho más fácil poder analizar el flujo de datos constante y en crecimiento que poseen, puesto que con los nuevos modelos como por ejemplo "Amazon Web Services" se ofrece toda un gama de servicios e infraestructura para que las empresas obtengan un mayor beneficio del análisis de sus datos, sin tener que montar la infraestructura propiamente, sino pagando por el alojamiento y consumo de los servicios en la nube.

6

Conclusiones

Las tecnologías que existen en la actualidad gracias al advenimiento de la era de la explosión de los datos "Big Data", ha provisto la habilidad de explorar los datos de tal manera que se puede tener conocimiento validado y en tiempo real del estado de las variables críticas del funcionamiento de un sistema. Toda la plataforma Big Data provee entonces poderosas herramientas para el análisis de datos hidroclimatológicos de una cuenca hídrica a través de la aplicación de las técnicas referenciadas en el presente trabajo y de la utilización de tecnologías como Hadoop, Spark, Bases de Datos NoSQL, Tableu, y lenguajes de programación como Python, R, D3.

Por otra parte y como se vió a lo largo de nuestra investigación un aspecto de suprema importancia es la presentación de los datos, puesto que escoger el modismo equivocado para representar la infomación, puede llevar a confusiones a la hora de interpretar los resultados del análisis sobre los datos. Es por ello, que los invitamos a estudiar en detalle el framework de visualización de datos que propuso Tamara Munzner con el fin de implementar las técnicas de análisis visual que a nuestro parecer llegan a ser de mucha más utilidad que otras técnicas de análisis.

A través de esta investigación se pudo notar cómo el uso de técnicas de análisis avanzadas dotan de capacidad a las máquinas y equipos computacionales en la realización de aquellas tareas que si no se automatizaran tomarían demasiado tiempo en completarse. Tal es el caso de los métodos vistos de Machine Learning, Minería de Datos, análisis de texto que se han vuelto de uso cotidiano en las labores de exploración de datos. A su vez, se pudo observar un caso de estudio puntual que nos enseña cómo de una manera sistemática y estructurada podemos definir nuestros propios modelos para aplicar esas técnicas de análisis Big Data en el estudio de las variables hidroclimatológicas de una cuenca hídrica, que en el caso de estudio descrito, se realizó sobre la cuenca que baña el departamento de Caldas, Colombia.

Bibliografía

[1] Hrushikesha Mohanty & Prachet Bhuyan. *Studies In Big Data*. Deepak Chenthati Editors, © Springer, India, 2015.

[2] Viktor Mayer-Schönberger & Kenneth Cukier. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt, Boston, New York, 2013. Disponible en: https://books.google.com.co/books?hl=es&lr=&id=uy4lh-WEhhIC&oi=fnd&pg=PP1&dq=types+data+big+data&ots=JshabnDPNM&sig=2PGBC1XUD5t_PQnIDTTy9vxuGr0#v=onepage&q=structured&f=false.

[3] Michael Schroeck [et al.]. *Analytics: el uso de big data en el mundo real: Cómo las empresas más innovadoras extraen valor de datos inciertos*. IBM® Institute for Business Value y la Escuela de Negocios Saïd en la Universidad de Oxford, 2013. Disponible en: http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf.

[4] Gil Press. *Big Data News: A Revolution Indeed*. Forbes, 2013. Disponible en: http://www.forbes.com/sites/gilpress/2013/06/18/big-data-news-a-revolution-indeed/#6c99afda7b9f.

[5] Winshuttle. «*A history of Big Data*». Disponible en: http://www.winshuttle.com/big-data-timeline/.

[6] IBM. *Relational Database*. Disponible en: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/reldb/.

[7] Martin Hilbert. *How to Measure "How Much Information"? Theoretical, Methodological, and Statistical Challenges for the Social Sciences*. International Journal of Communication, University of Southern California, 2012. Disponible en: http://ijoc.org/index.php/ijoc/article/view/1318/746.

[8] Michael Cox & David Ellsworth. *Application-Controlled Demand Paging for Out-of-Core Visualization*. NASA Ames Research Center, 1997. Disponible en: http://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf.

[9] Rajkumar Buyya & Rodrigo N. Calheiros & Amir Vahid Dastjerdi. *Big Data: Principles and Paradigms*. Elsevier Science, 2016. Disponible en: https://books.google.com.co/books?id=MfOeCwAAQBAJ.

[10] Kevin Ashton. *That 'Internet of Things' thing in the real world, things matter more than ideas*. RFID Journal, 2009. Disponible en: http://www.rfidjournal.com/articles/view?4986.

[11] Bret Swanson & George Gilder. «*Estimating the Exaflood*». Discovery Institute, 2008. Disponible en: http://www.discovery.org/a/4428.

[12] STAMFORD Conn. *Gartner EXP Worldwide Survey of More than 1,500 CIOs Shows IT Spending to Be Flat in 2009*. Gartner, Inc., 2009. Disponible en: http://www.gartner.com/newsroom/id/855612.

[13] Thor Olavsrud. *12 Big Data Predictions for 2014*. IDG Communications, 2013. Disponible en: http://www.cio.com/article/2369764/big-data/132163-12-Big-Data-Predictions-for-2014.html.

[14] Bernard Marr. *The Most Practical Big Data Use Cases Of 2016*. Forbes, 2016. Disponible en: http://www.forbes.com/sites/bernardmarr/2016/08/25/the-most-practical-big-data-use-cases-of-2016/#2b6355f07533.

[15] IBM. «*What is big data?*». Disponible en: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

[16] Andrew McAfee & Erik Brynjolfsson. *Big Data: the Management revolution*. Harvard business review, 2012.

[17] Amazon. *¿Qué son los big data?* Disponible en: https://aws.amazon.com/big-data/what-is-big-data/?nc1=h_ls.

[18] Wikipedia. «*Big data*». 2017. Disponible en: https://en.wikipedia.org/wiki/Big_data.

[19] GilPress. *What's The Big Data?* 2014. Disponible en: https://whatsthebigdata.com/2014/09/08/whats-the-big-data-12-definitions/.

[20] HowieT. *The Big Bang: How the Big Data Explosion Is Changing the World*. Microsoft, 2013. Disponible en: https://blogs.msdn.microsoft.com/microsoftenterpriseinsight/2013/04/15/the-big-bang-how-the-big-data-explosion-is-changing-the-world/.

[21] Jennifer Dutcher. «*What Is Big Data?*». Berkeley, 2014. Disponible en: https://datascience.berkeley.edu/what-is-big-data/.

[22] Doug Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Meta Group, 2001.

[23] IBM. *The Four V's of Big Data*. Infographics animations edition. Disponible en: http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

[24] IBM. *Extracting business value from the 4 V's of big data*. Infographics animations edition. Disponible en: http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

[25] Bernard Marr. *Big Data: The 5 Vs Everyone Must Know*. Linkedin, 2014. Disponible en: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know.

[26] Ben Larson. *Characteristics of Big Data*. educational research techniques, 2016. Disponible en: https://educationalresearchtechniques.wordpress.com/2016/05/02/characteristics-of-big-data/.

[27] Bill Vorhies. *How Many "V"s in Big Data – The Characteristics that Define Big Data*. Data Magnum, 2013.

[28] Disponible en: https://thesai.org/Downloads/Volume7No3/Paper_37-Extract_Five_Categories_CPIVW.pdf.

[29] Suhail Sami Owais & Nada Sael Hussein. *Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data*. IJACSA, vol. 7, no. 3 edition, 2016. Disponible en: http://www.datasciencecentral.com/profiles/blogs/top-10-list-the-v-s-of-big-data.

[30] Richard J. Self. *12 Vs of Big Data Governance*. SAS Institute Inc, University of Derby, 2014. Disponible en: http://computing.derby.ac.uk/c/wp-content/uploads/2012/11/Self_Richard_A2014.pdf.

[31] Dave Chaffey. *Digital Marketing Trends for 2017*. Smart insights [en línea] edition, 2017. Disponible en: http://www.smartinsights.

com/managing-digital-marketing/marketing-innovation/digital-marketing-trends-2016-2017/.

[32] Daniel Price. *Surprising facts and stats about the big data industry.* Cloudtweaks edition, 2015. Disponible en: https://cloudtweaks.com/2015/03/surprising-facts-and-stats-about-the-big-data-industry/.

[33] Susan Gunelius. *The Data Explosion in 2014 Minute by Minute – Infographic.* 2014. Disponible en: http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/.

[34] Viktor Mayer-Schönberger & Kenneth Cukier. *Big Data: la revolución de los datos masivos.* Madrid, turner publicaciones s.l edition, 2013.

[35] SAS. «*Big Data Analytics*». Disponible en: http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html.

[36] Amir Gandomi & Murtaza Haider. *Beyond the hype: Big data concepts, methods, and analytics.* Elsevier edition, 2015.

[37] Oracle. *Big Data Analytics - Advanced Analytics in Oracle Database.* 2013. Disponible en: http://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataanalyticswpoaa-1930891.pdf.

[38] Ingram Micro Advisor. *Four Types of Big Data Analytics Examples of Their Use.* Disponible en: www.ingrammicroadvisor.com.

[39] Halo Blog. *Descriptive, Predictive, and Prescriptive Analytics Explained.* 2016. Disponible en: halobi.com.

[40] Anwaar Ali & Junaid Qadir & Raihan ur Rasool & Arjuna Sathiaseelan & Andrej Zwitter & Jon Crowcroft. «*Big data for development: applications and techniques*». Biomed central edition, 2016.

[41] Frederick L. Oswald & Dan J. Putka. «*Statistical Methods for Big Data*». New York, in s. tonidandel & e. king & j. cortina edition, 2015.

[42] [Et.al] Torgyn Shaikhina. «*Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation*». Sciencedirect edition, 2017.

[43] Ned Horning. *Introduction to decision trees and random forest.* American museum of natural history's center for biodiversity and conservation edition, 2017.

[44] Andrea Trevino. *Introduction to K-means Clustering.* Datascience.com edition, 2016.

[45] Mark Brunelli. *Analysts: Data Visualization Tools Key to Big Data Analytics Success.* TechTarget, 2011. Disponible en: searchbussinessanalytics.techtarget.com.

[46] Vishal Gupta Poonam Vashisht. *Big Data Analytics Techniques: A survey.* IEEE, conference green computing and internet of things edition, 2015.

[47] Rebbeca Merrett. *5 Tools and Techniques for Text Analytics.* CIO, 2015. Disponible en: www.cio.com.au.

[48] IBM. *SPSS Text Analytics for Surveys.* Disponible en: www-03.ibm.com.

[49] IBM. *MapReduce.* Disponible en: https://www.ibm.com/analytics/us/en/technology/hadoop/mapreduce/.

[50] IBM. *Hadoop.* Disponible en: www-01.ibm.com.

[51] Maribel Tirados. *Apache Spark, la nueva estrella de Big Data.* 2014. Disponible en: www-01.ibm.com.

[52] IBM. *Hive.* Disponible en: www.ibm.com.

[53] M. Tim Jones. *Procese big data en tiempo real con Twitter Storm.* Disponible en: www.ibm.com.

[54] Ricardo Barranco Fragoso. *Análisis de Big Data con Apache Pig.* 2012. Disponible en: www.ibm.com.

[55] Ray Chiang & Dennis Dawson. *Untangling Apache Hadoop YARN, Part 1: Cluster and YARN Basics.* 2015. Disponible en: https://blog.cloudera.com/.

[56] Grant Ingersoll. *Introducing Apache Mahout.* 2009. Disponible en: www.ibm.com.

[57] Definición ABC. *Definición de Lluvia.* Disponible en: http://www.definicionabc.com/medio-ambiente/lluvia.php.

[58] Real Academia Española. *Pluviosidad*. Disponible en: http://dle.rae.es/?id=TSG4MHR.

[59] Ardikary Ariza & Alejandro Vargas & María Isabel González. *Pluviosidad*. Emaze, diapositivas edition. Disponible en: https://www.emaze.com/@ALQFLLIL/PLUVIOSIDAD.

[60] Alimentación y Medio Ambiente Ministerio de Agricultura y Pesca. *Sistema de Información del Anuario de Aforos*. Madrid. Disponible en: http://sig.mapama.es/redes-seguimiento/visor.html?herramienta=Aforos.

[61] *La dirección del viento*. 2004. Disponible en: http://webcache.googleusercontent.com/search?q=cache:http://www.oni.escuelas.edu.ar/2004/SAN_JUAN/676/eolica_y_molinos/capitulo_1/cap_1_2.htm.

[62] Wikipedia. *La Radiación Solar*. Es.wikipedia.org.

[63] Metas Metrólogos Asociados. *Presión atmosférica, presión barométrica y altitud. Conceptos y aplicaciones*. 2005.

[64] Wikipedia. *Ríos*. Es.wikipedia.org.

[65] Wikipedia. *Cuenca hidrográfica*. Es.wikipedia.org.

[66] Varvara Vetrova. *Selected Data Exploration Methods in Hydroclimatology*. 2016.

[67] M.karamouz & S. Nazif & M. Falahi. *Hidrology and Hydroclimatology: Principles and Applications*. Taylor & Francis, 2013. Disponible en: https://books.google.com.co/books?id=NnFLjSTCB6cC.

[68] A&V Ingeniería SAS. *Monitoreo Ambiental*. Disponible en: www.ayvingenieria.com.

[69] Lilian Morejón & Marina Vega & Antonio Escarré & José Peralta & Arely Quintero & Julio González. Análisis del balance hídrico en cuencas hidrográficas de la sierra de los Órganos. *Ingeniería Hidráulica y Ambiental*, 36, 2015.

[70] MSc Shaun Harrigan BA. *Exploring the Hydroclimatology of Floods: From Detection to Attribution*. 2016.

[71] Lorraine E. Flint & Alan L. Flint. *California Basin Characterization Model: A Dataset of Historical and Future Hydrologic Response to Climate Change: U.S. Geological Survey Data Release*. U.S. Geological Survey, 2017.

[72] Emilcy Juliana Hernández Leala. *Aplicación de técnicas de análisis de datos y administración de Big Data ambientales*. 2016. Disponible en: http://www.bdigital.unal.edu.co/54512/1/1090175695.2016.pdf.

[73] SAS. *Data Visualization. What it is and Why it Matters*. Disponible en: www.sas.com.

[74] Tamara Munzner. *Visualization Analytics and Design*. Taylor & Francis Group, 2015. Vista previa disponible en: www.crcpress.com.