



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Aplicación de técnicas de análisis de datos y administración de *Big Data* ambientales

Emilcy Juliana Hernández Leal

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ingeniería de la Organización
Medellín, Colombia
2016

Aplicación de técnicas de análisis de datos y administración de *Big Data* ambientales

Emilcy Juliana Hernández Leal

Tesis de investigación presentada como requisito parcial para optar al título de:

Magister en Ingeniería Administrativa

Director:

Ph.D. Néstor Darío Duque Méndez

Codirector:

Ph.D. Julián Moreno Cadavid

Línea de Investigación:

Procesamiento y análisis de datos, Inteligencia Artificial

Grupo de Investigación:

Grupo de Ambientes Inteligentes Adaptativos – GAIA

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ingeniería de la Organización

Medellín, Colombia

2016

*"Después de escalar una montaña muy alta,
descubrimos que hay muchas otras montañas
por escalar"*

Nelson Mandela

Agradecimientos

Agradezco a Dios, a mis padres y hermanito, ellos han sido mi apoyo toda la vida. A quien me ha dado un punto de vista objetivo cuando lo he necesitado y me ha acompañado en todo momento, Mao. A mis amigos del grupo GAIA, con quienes he compartido experiencias en la labor investigativa.

Durante este proceso han sido un pilar fundamental mi director Néstor Darío Duque Méndez y mi co-director Julián Moreno Cadavid, ellos guiaron mi proceso de investigación y más que docentes han sido unos excelentes orientadores y consejeros, han sabido llevarme al cumplimiento de los objetivos, tanto de la tesis como algunos personales; por ello les quiero expresar mi agradecimiento total. A los ingenieros del Instituto de Estudios Ambientales – IDEA, de la Universidad Nacional de Colombia en Manizales, sus aportes han sido muy valiosos para mí.

A los docentes e investigadores de la Universidad Politécnica de Valencia, con quienes realicé una pasantía investigativa de seis meses, agradezco por haberme acompañado en una de las experiencias más enriquecedoras de mi maestría, por haberme compartido parte de su tiempo y haber dado sus opiniones y valiosos aportes a mi trabajo, de ellos aprendí mucho.

Resumen

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

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

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

Palabras clave: Análisis de datos, *Big Data*, Datos Ambientales, Minería de Datos, Redes de Monitoreo Ambiental.

Abstract

The growth in the volume of data generated by different systems, as the measurement of daily activities, makes necessary to modify, optimize and develop data storage and processing methods and models able to supply the shortcomings presented the databases and KDD traditional processes. In this regard, *Big Data* analysis is an approach that includes several technologies associated with the storage, analysis, and visualization of big volumes of data obtained from several sources. Thus, *Big Data* analysis is a solution to the processing data issues that are not covered by the traditional solutions.

Moreover, the monitoring of ambient conditions, namely, climatic, meteorological and hydrometeorological constitute a data source growing exponentially because their measurements must be done in several spatial locations, with several capture strategies, and in many time instants.

Keywords: Analytical Data, *Big Data*, *Data Mining*, Environmental Data, Environmental Monitoring Networks.

Contenido

	Pág.
Resumen	IX
Abstract.....	X
Lista de figuras.....	XIV
Lista de tablas	XVI
Introducción	1
1. Presentación de la tesis.....	3
1.1 Motivación	3
1.2 Problemática.....	4
1.3 Preguntas de investigación	6
1.4 Objetivos.....	6
1.4.1 Objetivo general	6
1.4.2 Objetivos específicos	7
1.5 Alcance de la tesis.....	7
1.6 Metodología.....	8
1.7 Cumplimiento de los objetivos	10
1.8 Principales contribuciones logradas.....	12
1.9 Difusión de resultados	13
1.9.1 Artículos en revistas	13
1.9.2 Artículos en conferencias	13
1.9.3 Participación en proyectos	13
1.10 Organización del documento	14
2. Marco teórico.....	15
2.1 <i>Big Data</i>	15
2.1.1 Dimensiones de <i>Big Data</i>	17
2.1.2 Problemática en <i>Big Data</i>	17
2.2 Almacenamiento de datos	17
2.2.1 Bases de datos relacionales.....	18
2.2.2 Datawarehouse	18
2.2.3 Almacenamiento NoSQL.....	20
2.3 Técnicas de análisis de datos	21
2.3.1 Consultas o análisis relacional	22
2.3.2 Procesamiento analítico en línea OLAP	22
2.3.1 Minería de Datos.....	22
2.3.2 Machine Learning y Deep Learning.....	24

2.3.3 Tecnologías asociadas a <i>Big Data</i>	25
2.4 Componente ambiental	29
2.4.1 Redes y estaciones de monitoreo hidroclimatológico.....	30
2.4.2 Algunas variables hidrometeorológicas.....	31
3. Estado del arte	35
3.1 Primer grupo: tendencias en <i>Big Data</i>	35
3.2 Síntesis de los trabajos del primer grupo.....	47
3.3 Segundo grupo: almacenamiento y análisis de datos	48
3.4 Síntesis de los trabajos del segundo grupo	55
3.5 Tercer grupo: aplicación de técnicas para la predicción de comportamiento de variables climáticas.....	56
3.6 Síntesis del tercer grupo.....	70
4. Modelo propuesto.....	73
4.1 Identificación de las capas asociadas al modelo de administración de <i>Big Data</i>	73
4.1.1 Fuentes <i>Big Data</i>	74
4.1.2 Capa de almacenamiento	75
4.1.3 Capa de análisis	77
4.1.4 Capa de consumo.....	77
4.2 Tecnologías de apoyo para cada una de las capas del modelo.....	78
4.3 Modelo específico para la administración de <i>Big Data</i> ambiental	80
4.3.1 Fuentes de datos ambientales	81
4.3.2 Proceso de ETL	82
4.3.3 Almacenamiento	83
4.3.4 Análisis de datos.....	84
4.3.5 Presentación de datos	84
4.4 Conclusiones del capítulo.....	85
5. Validación del modelo. Caso de estudio	87
5.1 Implementación	87
5.2 Caso de estudio	88
5.3 Aplicación de proceso de ETL	90
5.4 Selección de la estrategia para la capa de almacenamiento	92
5.5 Método para la validación de la capa de análisis.....	93
5.5.1 Análisis predictivo usando Deep Learning	94
5.5.2 Análisis de clustering con <i>Mahout</i> y <i>Hadoop</i>	100
5.6 Presentación de resultados para la capa de consumo.....	107
5.7 Conclusiones del capítulo.....	110
6. Conclusiones y trabajos futuros.....	114
6.1 Conclusiones.....	114
6.2 Trabajos futuros	115
Bibliografía	119

Listas de figuras

	Pág.
Figura 2-1: Esquema general de los conceptos asociados a la tesis. Elaboración propia	15
Figura 2-2: Principales fuentes de datos para <i>Big Data</i> (George et al., 2014)	16
Figura 2-3: Marco de trabajo típico para la construcción de un Data Warehouse. Adaptado de (Han, Kamber, & Pei, 2011)	19
Figura 2-4: Modelos Datawarehouse. Adaptado de (Villanueva Chávez, 2011)	20
Figura 2-5: Modelos de almacenamiento NoSQL. Construido a partir de (Jaramillo Valbuena & Londoño, 2015).....	21
Figura 2-6: Tipos de OLAP. Elaboración propia	23
Figura 2-7: Esquema de los pasos del proceso de descubrimiento del conocimiento (KDD). Adaptado de (Han et al., 2011)	24
Figura 2-8: Pilares de un HDFS. Elaboración propia	26
Figura 2-9: Esquema general de los procesos <i>MapReduce</i> . Elaboración propia.....	27
Figura 3-1: Marco de trabajo para el procesamiento de <i>Big Data</i> . Tomado de (Wu et al., 2014)	37
Figura 3-2: Comparación entre HPCC y <i>Hadoop</i> . Tomado de (Sagiroglu & Sinanc, 2013)	41
Figura 3-3: Problemas abiertos en <i>Big Data</i> . Construido a partir de (Chen et al., 2014)..	42
Figura 3-4: Definiciones de <i>Big Data</i> basadas en una encuesta en línea realizada a 154 ejecutivos globales en abril de 2012. Tomado de (Gandomi & Haider, 2015)	46
Figura 3-5: Técnicas <i>Big Data</i> . Adaptado de (C. L. P. Chen & Zhang, 2014)	47
Figura 3-6: Integración de datos y procesos de minería usando datos ambientales distribuidos. Tomado de (Bartok et al., 2010).....	51
Figura 3-7: Arquitectura ADMIRE. Tomado de (Simo et al., 2011)	52
Figura 3-8: Arquitectura del sistema HaoLap. Tomado de (Song et al., 2015).....	54
Figura 3-9: Algoritmos y técnicas utilizadas. Tomado de (Duque Mendez et al., 2011)...	58
Figura 3-10: Esquema metodológico. Tomado de (Beltrán-Castro et al., 2013)	60
Figura 4-1: Capas del modelo genérico para la administración de <i>Big Data</i> . Elaboración propia.....	74
Figura 4-2: Tecnologías para almacenamiento. Elaboración propia.....	79
Figura 4-3: Modelo específico para la gestión de <i>Big Data</i> ambiental. Elaboración propia.	81
Figura 5-1: Aplicación del modelo específico al caso de estudio. Elaboración propia.....	88
Figura 5-2: Proceso de ETL aplicado a los datos del caso de estudio. Tomado de (Duque, Hernández, Pérez, Arroyave, & Espinosa, 2016)	91

Figura 5-3: Modelo E-R de la bodega de datos ambientales. Tomado de (Duque Méndez et al., 2015)	92
Figura 5-4: Flujo del modelo de análisis predictivo propuesto. (Hernández et al., 2016). 96	
Figura 5-5: Arquitectura: Autoencoder y Perceptrón Multicapa. (Hernández et al., 2016)97	
Figura 5-6: Resultados K-means con <i>Mahout</i> exportados a archivo plano	101
Figura 5-7: Resultados en consola para el algoritmo K-means con <i>Mahout</i>	102
Figura 5-8: Clústeres formados con un k=4. Elaboración propia.....	103
Figura 5-9: Resultados K-means con <i>Mahout</i> exportados a archivo plano para un k=5.103	
Figura 5-10: Clústeres formados con un k=5. Elaboración propia.....	104
Figura 5-11: Resultados en consola para el algoritmo Canopy con <i>Mahout</i>	105
Figura 5-12: Resultados del algoritmo Canopy con <i>Mahout</i> exportados a un archivo de texto plano.....	106
Figura 5-13: Clústeres formados con el algoritmo Canopy. Elaboración propia	106
Figura 5-14: Visualización de las capas en el aplicativo de estado del tiempo. Obtenido de http://idea.manizales.unal.edu.co/index.php/estado-tiempo-manizales	107
Figura 5-15: Presentación de datos detallados de cada estación en tiempo real. Obtenido de http://idea.manizales.unal.edu.co/index.php/estado-tiempo-manizales	108
Figura 5-16: Ejemplo de gráficas para la temperatura e indicador A25. Obtenido de http://idea.manizales.unal.edu.co/index.php/estado-tiempo-manizales	109
Figura 5-17: Presentación de datos históricos. Obtenido de cdiac.manizales.unal.edu.co	110

Lista de tablas

	Pág.
Tabla 1-1: Metodología: ciclos, objetivos y actividades	8
Tabla 2-1: Algunos algoritmos en <i>Mahout</i> . Tomado y adaptado de (Ingersoll, 2012)	28
Tabla 3-1: Comparativo de implementaciones de sistemas de almacenamiento. Tomado de (Jaramillo Valbuena & Londoño, 2015)	38
Tabla 3-2: Síntesis del segundo grupo	56
Tabla 3-3: Síntesis del tercer grupo	71
Tabla 4-1: Comparativo esquemas de almacenamiento. Adaptado de (Microsoft, 2014).76	76
Tabla 5-1: Redes de monitoreo del departamento de Caldas.....	89
Tabla 5-2: Información del conjunto de datos para el análisis predictivo. Elaboración propia.....	95
Tabla 5-3: Valores probados para cada parámetro. Elaboración propia.....	98
Tabla 5-4: Resultados de la predicción para cada método. Elaboración propia.....	99
Tabla 5-5: Resultados para las predicciones de Temperatura, Humedad y Presión. Elaboración propia	100

Introducción

El número de datos ambientales recolectados por estaciones de monitoreo climático tienen un significativo crecimiento, debido a que las mediciones son realizadas en tiempo real y entre menor sea el periodo de tiempo entre una medición y otra, mayor será la precisión de los indicadores climáticos calculados a partir de estos. Las redes de monitoreo hidrometeorológico miden variables como precipitación, temperatura del aire, humedad relativa, presión barométrica, dirección y velocidad del viento, nivel y caudal, entre otras; el número y tipo de variables depende de los diferentes sensores que sean instalados en cada estación y de las condiciones de su ubicación.

Lo anterior se convierte en una oportunidad para la aplicación de técnicas de análisis de datos y para el descubrimiento de conocimiento; sin embargo, se requiere de un esquema de administración de datos eficiente que incluya estrategias para la captura, entrega, almacenamiento y procesamiento que aseguren la calidad y consistencia de los datos (Duque-Méndez, Orozco-Alzate, & Vélez, 2014), y que a su vez permitan que el crecimiento del volumen de los datos no se convierta en un tema inmanejable y que las técnicas de análisis de datos se puedan aplicar, considerando diversos tipos y formatos de datos y la conexión a sus fuentes. *Big Data* es un término que incluye diferentes tecnologías asociadas a la administración de grandes volúmenes de datos provenientes de diferentes fuentes y que se generan con rapidez (Li, Jiang, Yang, & Cuzzocrea, 2015). Las tecnologías asociadas a *Big Data* han sido aplicadas a diferentes entornos como la seguridad y defensa, la social media, la medicina, la astronomía, entre otros, con buenos resultados; después de realizar la revisión bibliográfica se vislumbra que una aplicación concreta para datos ambientales podría obtener resultados interesantes y positivos.

Teniendo en cuenta lo anterior, se realizó el planteamiento de esta investigación, como tesis de maestría con el objetivo de atender a la problemática y las oportunidades

enunciadas, por medio de la formulación de un modelo para la administración de datos ambientales, que incluya la utilización de tecnologías asociadas a *Big Data* y técnicas de análisis de datos basadas en inteligencia artificial.

1. Presentación de la tesis

En este capítulo se hace la presentación de aspectos generales de la tesis, los cuales son la base para el desarrollo posterior. Se introduce a continuación la motivación, área problema, preguntas de investigación, objetivos, alcance, metodología y logros alcanzados. Finalmente se describe la organización del documento.

1.1 Motivación

Dentro de las líneas de investigación del Grupo de Investigación en Ambientes Inteligentes Adaptativos - GAIA, al que pertenece la investigadora y uno de sus asesores, se encuentra el procesamiento y análisis de datos; dicha línea se enfoca en el desarrollo de técnicas de almacenamiento y análisis de datos para permitir el aprovechamiento de grandes volúmenes de datos, con la extracción de información relevante y útil para la toma de decisiones. A partir de lo anterior y de la estrecha relación del Grupo GAIA con el Instituto de Estudios Ambientales – IDEA – de la Universidad Nacional de Colombia Sede Manizales, quienes administran estaciones hidrometeorológicas y cuentan con bases de datos ambientales de largos períodos de tiempo, nace el interés de abordar la temática.

Los aspectos ya mencionados y la revisión bibliográfica inicial, inducen a explorar en otros modelos de administración y en técnicas para análisis de este tipo de datos. Esta tesis busca atender la problemática asociada al creciente volumen de datos ambientales, provenientes de diferentes variables de tipo climático (meteorológicas e hidrometeorológicas) y a la oportunidad que se presenta de extraer conocimiento de estos datos que permita además del análisis y comprensión de los mismos, la toma de decisiones y la posible predicción de eventos asociados a dichas variables.

Así mismo, el plan de estudios cursado, Maestría en Ingeniería Administrativa, propende al desarrollo de capacidades y conocimientos en administración aplicada a diferentes campos de acción que han emergido en la sociedad actual, esto incluye el desarrollo de

investigación aplicada y en lo posible la innovación por medio de avances tecnológicos que lleve a la concepción de soluciones a problemas de tipo administrativo. Siendo lo anterior también, parte de los propósitos que se plantearon dentro de esta tesis.

1.2 Problemática

Los grandes volúmenes de datos ambientales (meteorológicos, climáticos e hidrometeorológicos) generados por estaciones de monitoreo ambiental, son una gran oportunidad para describir las variables en el tiempo según la ubicación de las tomas de datos, registrar los cambios ocurridos, conocer el comportamiento de las diferentes medidas, descubrir relaciones entre los datos y patrones en la dinámica de los fenómenos; pero el gran volumen de datos obtenidos exige que se usen herramientas informáticas para su almacenamiento y gestión que permitan tener un mecanismo eficiente de análisis de datos y extracción de información y conocimiento relevante para la toma de decisiones y la generación de alertas tempranas por medio de la predicción de comportamientos. Es decir, es necesario contar con un modelo de gestión de datos para el análisis y extracción de información concreta que ayude, tanto a la toma de decisiones como a la construcción de mecanismos de alerta temprana, basados en predicción de eventos a partir del estudio de los datos históricos.

Siguiendo los planteamientos anteriores, se tiene que el volumen creciente de datos ambientales con diferentes variables, escalas y nivel de detalle se convierte en un problema y en una oportunidad que requiere la definición e implementación de técnicas de análisis de datos que permitan aprovechar dichos datos disponibles para extraer conocimiento y que sirvan de apoyo a la toma de decisiones que además puedan generalizarse a otros tipos de datos y otros espacios geográficos. El uso de técnicas de almacenamiento como *data warehouse* y de análisis como OLAP o de minería de datos han permitido avanzar en el tema y se reportan resultados importantes desde diferentes comunidades (Duque-Méndez et al., 2014). En particular en *Data Mining* es usual que las técnicas descriptivas o predictivas utilizadas se apliquen sobre *dataset* extraídos de los datos por conveniencia o necesidad del trabajo en concreto (Larose & Larose, 2014). Con respecto a las técnicas de análisis multidimensionales (OLAP) estas se ven como un enfoque llamativo para lograr el cruce de variables, pero los costos de las consultas se

expresan como dificultades para su implementación práctica y generalizada (Chaudhuri & Dayal, 1997), en particular en ambientes disponibles para usuarios diversos. Por su parte, el almacenamiento y actualización de los datos registrados de estaciones de monitoreo con lapsos de tiempo pequeños (minutos) implica un proceso de limpieza, filtrado y transformación (ETL) con problemas ampliamente conocidos en la comunidad (Bustamante Martínez, Galvis Lista, & Gómez Flórez, 2013), y por otro lado el tipo y la estructura del esquema de almacenamiento se ve reflejado en la eficiencia de los modelos y la disponibilidad para la aplicación de técnicas de análisis de datos.

La tendencia de *Big Data* incluye más que la generación de nuevas tecnologías, estudios y profesiones nacientes como los analistas de datos, es un fenómeno que afecta innumerables aspectos en la sociedad. “En un mundo en el que hemos llegado al zettabyte de información digitalizada, tanto nuestra relación con el procesamiento de datos como la forma de mostrarlos al usuario deben cambiar. Ya no podemos interpretar la información sin la mediación del ordenador; necesitamos el software para procesar pero también para dar forma a un volumen de información que escapa a nuestra capacidad cognitiva” (Adell & Guersenzvaig, 2013).

Big Data se ha convertido en una expresión muy utilizada para denotar un campo de aplicación en el que bases de datos muy grandes, generalmente no estructuradas y no relacionales pueden ser analizadas, administradas, organizadas y finalmente utilizadas para dar soporte a los negocios. La importancia de esta tendencia ha abarcado tanto el campo de la industria, como la computación, la economía, la academia y la investigación. *Big Data* plantea retos considerables para los investigadores, que incluyen aspectos funcionales y no funcionales de los datos, desafíos como la escalabilidad de la computación y del almacenamiento de los datos, las tecnologías de consulta y procesamiento de los datos, las técnicas y herramientas para la administración y planificación y finalmente la tolerancia a fallos (Barbierato, Gribaudo, & Iacono, 2014).

Cómo se aprecia el enfoque y las tecnologías de *Big Data*, pueden ofrecer alternativas combinadas armónicamente con las tradicionales, para la gestión de datos ambientales, desde su recuperación hasta la extracción de conocimiento, pero dado lo reciente de su auge exige un proceso riguroso de análisis de los diferentes componentes y la adaptación y aprovechamiento específico a los datos relacionados con esta tesis.

1.3 Preguntas de investigación

Una vez planteado el problema de investigación y después de haber realizado la revisión del estado del arte, han surgido una serie de preguntas de investigación, las cuales se plantean a continuación:

- ¿Es posible aplicar el enfoque *Big Data* a los datos ambientales hidroclimatológicos y resolver problemas presentes en la aplicación de otros enfoques?
- ¿Es posible desarrollar un modelo para la aplicación de tecnologías asociadas a *Big Data* en grandes volúmenes de datos ambientales?
- ¿Qué tecnologías asociadas a *Big Data* se pueden adaptar mejor para el tratamiento de grandes volúmenes de datos ambientales?
- ¿De qué manera se puede modelar un sistema de predicción de eventos asociados al clima con base en datos ambientales e incorporando técnicas de inteligencia artificial y *Big Data*?

1.4 Objetivos

Los objetivos de esta tesis surgen a partir de la problemática planteada y de las preguntas de investigación formuladas. A continuación se presenta el objetivo general y se hace su desglose en los objetivos específicos.

1.4.1 Objetivo general

Proponer y validar un modelo de administración y análisis de datos ambientales (meteorológicos e hidrometeorológicos) generados por estaciones de monitoreo ambiental, a través de la utilización de *Big Data* y de técnicas de inteligencia artificial, para mejorar la

toma de decisiones y la generación de alertas tempranas por medio de la predicción de comportamientos.

1.4.2 Objetivos específicos

- Identificar las fuentes de datos ambientales (meteorológicos e hidrometeorológicos) generados por estaciones de monitoreo ambiental, tecnologías asociadas a *Big Data* y algunas técnicas de inteligencia artificial, para la obtención de información que será utilizada como base para la construcción del modelo de almacenamiento, administración y análisis de datos.
- Evaluar la utilización de tecnologías asociadas a *Big Data* para la construcción de modelos de administración y análisis de datos ambientales.
- Caracterizar y seleccionar los elementos de *Big Data* a utilizar y hacer el diseño del modelo de almacenamiento, administración y análisis de datos que permita la extracción de información relevante para mejorar la toma de decisiones y la generación de alertas tempranas.
- Validar el modelo de aplicación de técnicas de análisis y tecnologías de *Big Data* propuesto en un caso de estudio en datos ambientales.

1.5 Alcance de la tesis

En el desarrollo de esta tesis se considera los siguientes aspectos como delimitantes del alcance de la misma:

- El modelo de administración y análisis de datos ambientales abarcará la utilización de algunas tecnologías asociadas a *Big Data* y de las técnicas de inteligencia artificial que se consideran en el marco conceptual.

- Las predicciones generadas a partir de los datos del caso de estudio no llegarán a ser probadas ni se garantiza que se ajusten a eventos futuros. La validación y ajustes a estas predicciones hace parte de un trabajo futuro.
- Todos los análisis de información generada de la aplicación del modelo en casos de estudio queda sujeta a la revisión de expertos en el campo ambiental, específicamente en meteorología y en gestión de riesgos; ya que una incorrecta interpretación de esta información puede generar consecuencias significativas debido a la complejidad de los fenómenos naturales monitoreados y a los eventos asociados a estos.

1.6 Metodología

Para la ejecución de la tesis se plantean cuatro ciclos, cada ciclo está asociado al desarrollo de uno de los objetivos específicos presentados en la sección 1.4, estos son a su vez caracterizados en una serie de actividades. Los ciclos y actividades de la metodología se resumen en la tabla 1-1.

Tabla 1-1: Metodología: ciclos, objetivos y actividades

CICLO	OBJETIVO ASOCIADO	ACTIVIDADES
Ciclo 1: Revisión y refinamiento del marco conceptual y preparación de fuentes de datos para el modelo.	Objetivo 1: Identificar las fuentes de datos ambientales (meteorológicos e hidrometeorológicos) generados por estaciones de monitoreo ambiental, tecnologías asociadas a <i>Big Data</i> y algunas técnicas de inteligencia artificial, para la obtención de información que será utilizada como base para la construcción del modelo de almacenamiento, administración y análisis de datos.	<ul style="list-style-type: none"> • Revisión de bibliografía sobre técnicas de análisis de datos que se apoyen en inteligencia artificial y tecnologías asociadas a <i>Big Data</i>; teniendo en cuenta arquitecturas, métodos para el diseño y herramientas. • Revisión y caracterización de las fuentes de datos ambientales que alimentarán el modelo; haciendo énfasis en la naturaleza de los datos

CICLO	OBJETIVO ASOCIADO	ACTIVIDADES
		<p>y el tratamiento necesario antes de iniciar el análisis.</p> <ul style="list-style-type: none"> • Revisión y análisis de modelos que empleen <i>Big Data</i> para la administración de datos tanto ambientales como de otra naturaleza.
Ciclo 2: Estudio y comparación de tecnologías para el almacenamiento, administración y análisis de grandes volúmenes de datos.	Objetivo 2: Evaluar la utilización de tecnologías asociadas a <i>Big Data</i> para la construcción de modelos de administración y análisis de datos ambientales.	<ul style="list-style-type: none"> • Análisis y cotejo de las técnicas y tecnologías utilizadas para la construcción de modelos de almacenamiento y administración de grandes volúmenes de datos. • Estructuración de las funcionalidades y principales características que se incluirán en el modelo.
Ciclo 3: Diseño del modelo para administración y análisis de datos ambientales	Objetivo 3: Caracterizar y seleccionar los elementos de <i>Big Data</i> a utilizar y hacer el diseño del modelo de almacenamiento, administración y análisis de datos que permita la extracción de información relevante para mejorar la toma de decisiones y la generación de alertas tempranas.	<ul style="list-style-type: none"> • Conceptualización del modelo y análisis de los requerimientos para el tratamiento de datos ambientales. • Diseño de los componentes del modelo y de la arquitectura que se asociará a este. • Acoplamiento de los componentes del modelo.
Ciclo 4: Construcción y validación del prototipo.	Objetivo 4: Validar el modelo de aplicación de técnicas de análisis y tecnologías de <i>Big Data</i> propuesto en un caso de estudio en datos ambientales.	<ul style="list-style-type: none"> • Desarrollo, implementación e integración del prototipo del modelo. • Refinamiento de los datos ambientales elegidos como caso de estudio para la validación del prototipo del modelo. • Ejecución del prototipo para el caso de estudio.

CICLO	OBJETIVO ASOCIADO	ACTIVIDADES
		<ul style="list-style-type: none"> • Evaluación preliminar de los resultados obtenidos después de la puesta en marcha del prototipo.

1.7 Cumplimiento de los objetivos

Cada uno de los objetivos de la tesis fue abordado mediante la ejecución de sus respectivas actividades; a continuación, se describe de manera sintética la estrategia usada para dar cumplimiento a cada uno de estos.

Objetivo 1: *Identificar las fuentes de datos ambientales (meteorológicos e hidrometeorológicos) generados por estaciones de monitoreo ambiental, tecnologías asociadas a Big Data y algunas técnicas de inteligencia artificial, para la obtención de información que será utilizada como base para la construcción del modelo de almacenamiento, administración y análisis de datos.*

Este objetivo estaba asociado al ciclo 1 y su desarrollo se realizó por medio de una revisión bibliográfica que permitió identificar los principales conceptos relacionados con técnicas de análisis de datos apoyadas en la inteligencia artificial y tecnologías *Big Data*, esto se estructuró en el marco teórico que se presenta en el capítulo 2. También se revisaron algunos trabajos previos desarrollados en el área problemática, se identificaron sus fortalezas y limitaciones, construyendo con esto el estado del arte que se muestra en el capítulo 3 y que abarca tres grupos. El primer grupo de trabajos revisados comprende temas relacionados con *Big Data*, aspectos conceptuales, avances y retos en el área; el segundo grupo, comprende artículos que presentan el diseño y/o implementación de modelos para el almacenamiento de grandes volúmenes de datos y procesos de análisis sobre estos; finalmente, el tercer grupo se concentra en la aplicación de técnicas de inteligencia artificial, para la realización de predicciones de variables ambientales.

Objetivo 2: *Evaluar la utilización de tecnologías asociadas a Big Data para la construcción de modelos de administración y análisis de datos ambientales.*

Para afrontar este objetivo se realizó una búsqueda de las tecnologías asociadas a *Big Data* y se identificó que las estrategias para el almacenamiento y administración de datos en general son bastante variadas, dependen en gran medida del dominio en el que se desee realizar la implementación. En el caso particular de los datos ambientales y específicamente las mediciones hidrometeorológicas, éstas presentan una particularidad y es que se producen en tiempo real y como tal deben ser capturadas, almacenadas y procesadas teniendo en cuenta esta propiedad. Además de los datos que son generados para las estaciones de monitoreo ambiental, también se pueden tener otras fuentes como lo son los datos generados por medio de radares o satélites. En esta medida, se encontró que para el tratamiento de estos tipos de datos una buena alternativa de almacenamiento son las bodegas de datos, ya que estas permiten manejar diferentes dimensiones y se adaptan al manejo de un volumen considerable de datos y al crecimiento de estos. En cuanto a la utilización de tecnologías *Big Data* para el análisis de datos ambientales se presentan una variedad de soluciones con potencial para ser aplicadas, estas tecnologías se caracterizan por trabajar siguiendo el paradigma de divide y vencerás, razón por la cual requieren de buenos recursos para la distribución de tareas y para lograr ser eficientes. Este objetivo se ve abarcó igualmente a partir de la revisión de literatura.

Objetivo 3: Caracterizar y seleccionar los elementos de *Big Data* a utilizar y hacer el diseño del modelo de almacenamiento, administración y análisis de datos que permita la extracción de información relevante para mejorar la toma de decisiones y la generación de alertas tempranas.

Este objetivo se asocia al ciclo principal de la tesis, y se divide en dos grandes partes, en primer lugar el planteamiento y diseño conceptual del modelo genérico por capas para el tratamiento de datos masivos y en segundo lugar la inclusión el planteamiento de un modelo específico para datos ambientales, incluyendo tecnologías de análisis tradicionales y otras asociadas a *Big Data*. Para dar cumplimiento a este objetivo se partió de las oportunidades y dificultades encontradas en el estado del arte, a su vez se consideró las tecnologías revisadas en el marco teórico. Cada una de las capas del modelo genérico y específico se describe y presenta en el capítulo 4.

Objetivo 4: Validar el modelo de aplicación de técnicas de análisis y tecnologías de *Big Data* propuesto en un caso de estudio en datos ambientales

Para el ciclo que incluye este objetivo se tomó como caso de estudio, los datos hidrometeorológicos recolectados por el Instituto de estudios Ambientales – IDEA – de la Universidad Nacional de Colombia Sede Manizales y se estructuró un esquema particular para su tratamiento, también se aplicaron algunos análisis y se generaron unas predicciones de comportamientos futuros, las cuales quedan sujetas a validación.

1.8 Principales contribuciones logradas

En el marco de esta tesis, se realizó una investigación que se orientó hacia el fortalecimiento de una estrategia para el tratamiento de datos ambientales, incluyendo desde su almacenamiento hasta la extracción de información relevante y posibles predicciones. Pensando en esto se propuso un modelo de almacenamiento, administración y análisis de datos genérico y uno específico para datos ambientales con la posibilidad de incluir tecnologías asociadas a *Big Data*.

El modelo específico que se propone trata de hacer frente a algunas de las problemáticas que fueron identificadas en la revisión del estado del arte y que hacen parte del dominio específico de los datos ambientales. El modelo genérico incluye tres capas: una capa de almacenamiento, otra de análisis y una final de consumo. En cada capa se pueden incluir tecnologías *Big Data* o tradicionales buscando obtener la solución que mejor se acople a la gestión de los datos.

Las principales contribuciones logradas con esta tesis se pueden resumir en los siguientes aspectos:

- ✓ Desde una perspectiva conceptual se planteó un modelo multicapa para la gestión de datos que abarca almacenamiento, análisis y consumo de los mismos, pensando en la posibilidad de acoplar tecnologías *Big Data* en una o más de estas capas; a su vez, se detalló este modelo en otro específico y adaptado al dominio de datos ambientales.
- ✓ Desde la perspectiva de investigación aplicada (investigación y desarrollo) se trabajó en varios aspectos del almacenamiento, limpieza, análisis de los datos y predicción de comportamientos, aplicando esto al caso de estudio seleccionado.

1.9 Difusión de resultados

Como parte del proceso investigativo se realizó difusión de algunos de los resultados alcanzados. A continuación se presentan los trabajos publicados y en proceso de revisión que han surgido del desarrollo de esta Tesis y por el trabajo colaborativo con el director y codirector de la misma, algunos compañeros del grupo GAIA e integrantes del grupo GTI-IA de la Universidad Politécnica de Valencia.

1.9.1 Artículos en revistas

Hernández, Emilcy; Duque, Néstor; Moreno, Julián. "Generación de pronósticos para la precipitación diaria en una serie de tiempo de datos meteorológicos". 2016. Tunja, Colombia. Revista Ingenio Magno ISSN: 2145-9282. Editorial Universidad Santo Tomás. Volumen 7, Número 1, p.p144 – 155.

Duque, Néstor; Hernández, Emilcy; Pérez, Ángela; Arroyave, Adrián; Espinosa, Daniel. "Modelo para el proceso de extracción, transformación y carga en bodegas de datos. Una aplicación con datos ambientales". Julio – Diciembre 2016. Bogotá, Colombia. Revista Ciencia e Ingeniería Neogranadina ISSN: 0124-8170. Editorial Universidad Militar Nueva Granada. Volumen 26, Número 2, p.p 2 – 17.

Hernández, Emilcy; Duque, Néstor; Moreno, Julián. "*Big Data*: una revisión de investigaciones y casos de aplicación". Revista Tecno-Lógicas. Enviado, en revisión.

1.9.2 Artículos en conferencias

E. Hernández, V. Sánchez-Anguix, V. Julian, J. Palanca and N. Duque. Rainfall prediction: A Deep Learning approach. In Proceedings 11th International Conference on Hybrid Artificial Intelligence Systems. pp 151-163, 2016

1.9.3 Participación en proyectos

Proyecto de Investigación: Fortalecimiento de capacidades conjuntas para el procesamiento y análisis de información ambiental. Financiado por la Universidad Nacional de Colombia.

Proyecto de Extensión: Aunar esfuerzos para mejorar la gestión del riesgo mediante la incorporación del riesgo en la planificación y la toma de conciencia en el municipio de Manizales fase 1.

1.10 Organización del documento

El documento se estructura de la siguiente manera, en el capítulo 2 se describe el marco teórico que comprende los principales temas asociados a esta investigación; en el capítulo 3 se hace una recopilación de algunos trabajos relacionados con la problemática abordada, revisando sus fortalezas y limitaciones. En el capítulo 4 el modelo propuesto es presentado; el capítulo 5 contiene la descripción de los datos del caso de estudio, así como también la aplicación y validación del modelo. Finalmente las conclusiones y trabajos futuros son planteados en el capítulo 6.

2. Marco teórico

Por medio de una revisión preliminar se determinaron los principales conceptos asociados al tema planteado en esta propuesta, los cuales se presentan a continuación. La Figura 2-1 recoge algunos de los conceptos en que se basa esta tesis.

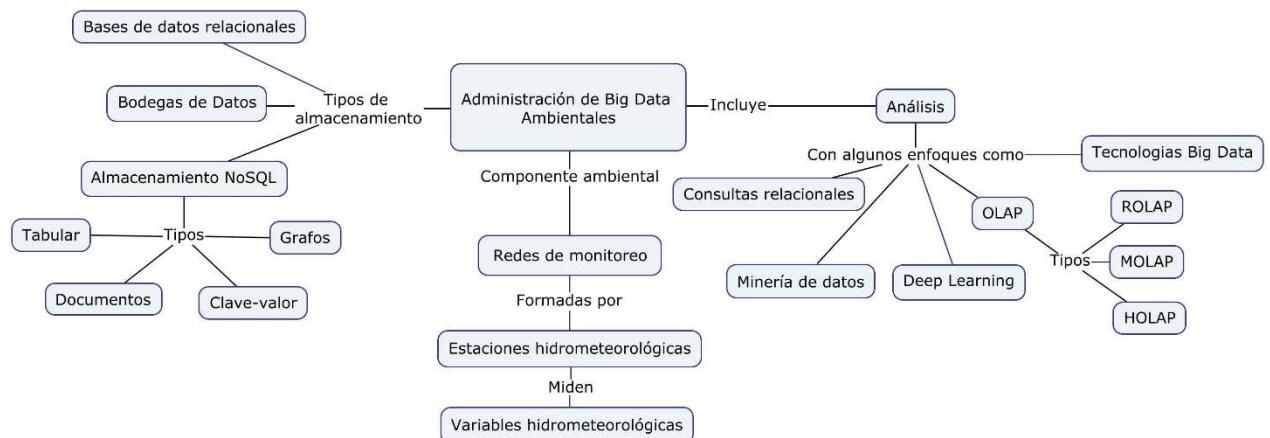


Figura 2-1: Esquema general de los conceptos asociados a la tesis. Elaboración propia

2.1 Big Data

El término *Big Data* es una expresión usada para hacer referencia a “grandes conjuntos de datos digitales que requieren de sistemas informáticos para su captura, almacenamiento, búsqueda, manipulación y visualización” (Adell & Guersenzvaig, 2013). Este crecimiento está asociado al aumento en la capacidad de los computadores, la instalación de sensores de diferente índole y las interacciones de los usuarios con sistemas de información e incluso con la web y las redes sociales. Según (George et al., 2014), *Big Data* se puede ver como un contenedor de diferentes tipos de datos, y los agrupan en cinco fuentes principales, las cuales se esquematizan en la Figura 2-2.

Los datos públicos son datos que típicamente son producidos por el gobierno, las organizaciones gubernamentales y las comunidades locales, una característica de estos datos es que pueden ser accedidos sin restricciones. Por su parte, los datos privados son producidos por compañías particulares, organizaciones sin ánimo de lucro e individuos y requieren políticas de protección de datos, y parte de ellos no pueden ser fácilmente accedidos por el público; por ejemplo las transacciones de los consumidores de un establecimiento, uso de teléfonos móviles, la navegación por ciertas partes de la web, etc. Los "data exhaust" se refieren a datos del ambiente que son recolectados pasivamente, estos datos pueden ser recolectados para diferentes propósitos pero también pueden ser combinados con otras fuentes de datos y crear nuevos valores, ejemplos de estos pueden ser las búsquedas en internet. Los datos de comunidades son datos no estructurados, especialmente texto, recolectados de la dinámica de las diferentes redes y tendencias sociales. Por último, los datos auto-cuantificados son un tipo de datos que se revelan por los individuos a través de la cuantificación de acciones personales diarias, que incluyen comportamientos y preferencias; un ejemplo son los datos tomados por una banda para ejercicio físico y subidos al celular por medio de una aplicación móvil (George, Haas, & Pentland, 2014).

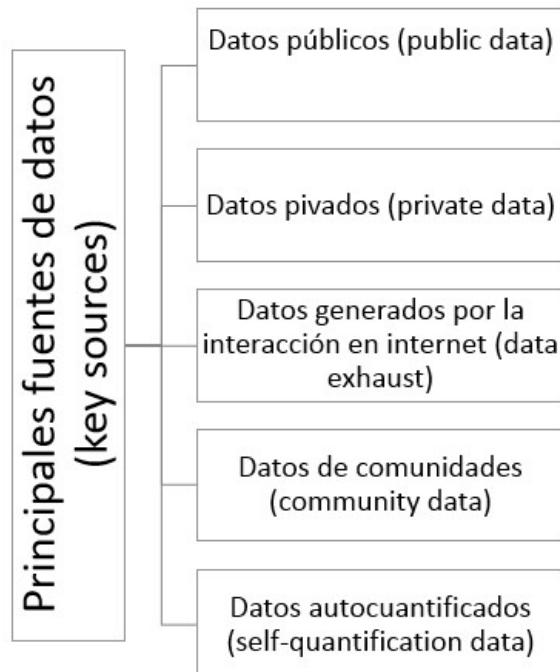


Figura 2-2: Principales fuentes de datos para *Big Data* (George et al., 2014)

2.1.1 Dimensiones de *Big Data*

Cuando se habla de *Big Data*, el término se suele asociar al tratamiento de grandes volúmenes de datos, sin embargo, *Big Data* es un enfoque que abarca otras dimensiones. A continuación se nombran y describen brevemente. Para algunos autores, *Big Data* ha pasado a ser el enfoque de las 4Vs (Talburt & Zhou, 2015).

- Volumen: acumulación y constante aumento de datos
- Velocidad: necesaria para generar, obtener y procesar los datos
- Variedad: de origen, fuente y formato de los datos
- Veracidad: fiabilidad de los datos

2.1.2 Problemática en *Big Data*

Dentro de los grandes retos que se presenta en la actualidad para los seres humanos es la necesidad de diseñar y construir nuevos modelos y metodologías para el almacenamiento y tratamiento de este creciente número de datos. En esta medida, *Big Data* se ha convertido en un potencial candidato para dar solución parcial o total a este reto y permitir una transformación a los datos para poder convertirlos en conclusiones y sacar provecho de los históricos acumulados durante años (Adell & Guersenzvaig, 2013).

Otro de los mayores cambios que se espera introducir con la utilización de *Big Data* es el aumento en la capacidad y velocidad para la toma de decisiones, lo cual sería un avance significativo para las organizaciones, ya que diferentes procesos como el desarrollo de nuevos productos, la administración de los clientes, la cadena de suministro, entre otras, podrían optimizarse e incluso llegar a un ideal de toma de decisiones en tiempo real (Galbraith, 2014). Esto mismo aplica a otros sectores como la medicina, la seguridad, la agricultura, la climatología, el transporte, los medios de comunicación, entre otros; los cuales se verían beneficiados al contar con este tipo de ventajas.

2.2 Almacenamiento de datos

Uno de los aspectos importantes a considerar en un sistema para la administración de un gran volumen de datos es el almacenamiento; dado que de esto dependen muchos factores como acceso, disponibilidad, escalabilidad, facilidad de recuperación,

estructuración de consultas, tiempos de respuesta, entre otras. Existen diferentes tecnologías para el almacenamiento de grandes volúmenes de datos, las cuales presentan sus ventajas y limitaciones; van desde los modelos más tradicionales (relacionales) hasta las nuevas tendencias que incluyen la posibilidad de almacenar datos no estructurados o semiestructurados. A continuación se presentan algunos de los modelos más representativos.

2.2.1 Bases de datos relacionales

Se denomina bases de datos relacionales a los esquemas de almacenamiento que cumplen con un modelo relacional, este tipo de bases de datos permiten las conexiones o relaciones entre los registros que están contenidos en tablas (Silberschatz et al., 2002). Tienen una serie de características particulares, entre ellas están: se componen de varias tablas y relaciones, no se pueden construir dos tablas con el mismo nombre, se denomina clave primaria al registro que identifica a la tabla y deben cumplir con el principio de integridad de datos. Para la manipulación de los datos contenidos en estas bases de datos se utiliza el álgebra relacional y el cálculo relacional. Como lenguaje común para realizar consultas a las bases de datos relacionales se presenta el SQL (Structured Query Language), este se encuentra implementado por los motores o sistema de gestión de bases de datos relacionales.

2.2.2 Datawarehouse

Un Datawarehouse o bodega de datos es un repositorio de información histórica recolectada de múltiples fuentes, unificada bajo un esquema y que usualmente se encuentra en un mismo lugar. Las bodegas de datos están construidas por medio de un proceso de limpieza, integración, transformación, carga y actualización periódica de datos (Han, Kamber, & Pei, 2011). En la Figura 2-3 se presenta un esquema general para la construcción de un Datawarehouse.

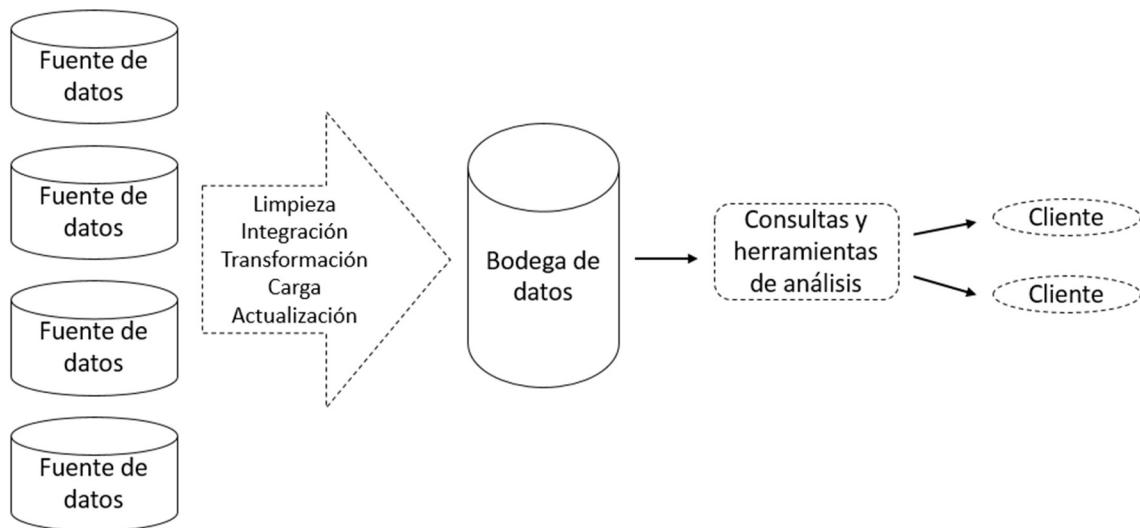


Figura 2-3: Marco de trabajo típico para la construcción de un Data Warehouse.
Adaptado de (Han, Kamber, & Pei, 2011)

Las arquitecturas utilizadas en los datawarehouse son relacional y multidimensional, la segunda presenta dos estructuras principales, estructura en estrella y estructura copo de nieve. En la arquitectura multidimensional se tienen una serie de atributos agrupados en unas dimensiones que a su vez hacen parte de un esquema (Han et al., 2011). Las bodegas de datos son utilizadas en diversos campos y organizaciones, ya que el crecimiento del volumen de datos es generalizado y se ha convertido en una tendencia desde hace tiempo atrás. Existen varios aspectos fundamentales a la hora de caracterizar un datawarehouse, como son la orientación a un tema específico, la integración, la no volatilidad y la variación en el tiempo. El hecho de que un datawarehouse sea integrado, implica que va a ser alimentado con datos provenientes de diferentes fuentes, los cuales deberán ser limpiados y estructurados bajo un esquema. Ahora bien, cuando se habla de que un datawarehouse debe ser no volátil, implica que contrario a como pasa en los sistemas transaccionales tradicionales (donde se inserta y modifica información de forma constante), en un datawarehouse los datos se cargan y acceden generalmente de forma masiva sin ser modificados (Shi, Lee, Duan, & Wu, 2001).

Los datawarehouse pueden estar diseñados bajo modelos multidimensionales, en los cuales se tiene la presencia de una tabla de hechos que es rodeada de dimensiones, las principales arquitecturas utilizadas son la estructura en estrella y la estructura copo de

nieve. En la Figura 2-4 se presenta un esquema de estas estructuras. La estructura en estrella es el modelo multidimensional clásico, con una única tabla de hechos rodeada de dos o más tablas de dimensiones. Por su parte, el copo de nieve es una variante que presenta varias tablas de hechos que comparten algunas tablas de dimensiones entre sí (Villanueva Chávez, 2011).

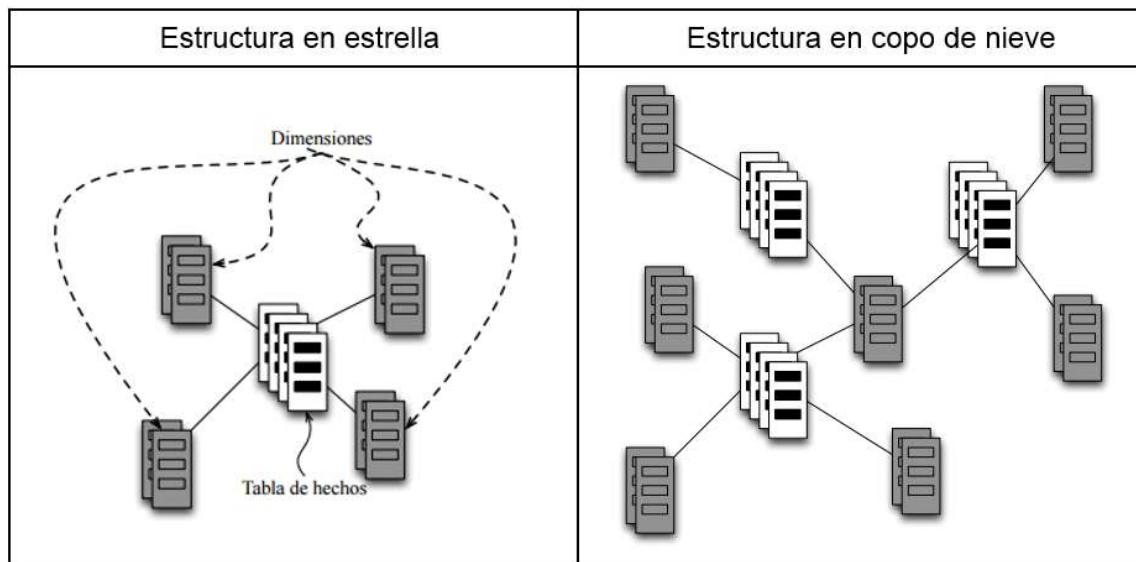


Figura 2-4: Modelos Datawarehouse. Adaptado de (Villanueva Chávez, 2011)

2.2.3 Almacenamiento NoSQL

Una alternativa que ha surgido para el almacenamiento de grandes volúmenes de datos y que contempla datos no estructurados o semiestructurados es el movimiento NoSQL (Not Only SQL), el cual propone el uso de sistemas que permitan gestionar estos grandes volúmenes de datos de forma eficiente y económica, cumpliendo con características como: capacidad de escalar horizontalmente y de replicar y distribuir datos a lo largo de muchos servidores; tener una interfaz simple de nivel de llamada; presentar un modelo de concurrencia débil; hacer un uso eficiente de índices distribuidos y memoria RAM para el almacenamiento de los datos y finalmente, dar la posibilidad de añadir de forma dinámica nuevos atributos a los registros de datos (Jaramillo Valbuena & Londoño, 2015). Los cuatro modelos de almacenamiento NoSQL más usados son presentados en la Figura 2-5.

Frente a los sistemas de bases de datos relacionales tradicionales, los sistemas NoSQL son recomendados cuando se requiere atender a millones de usuarios sin perder el rendimiento, como es el caso de las redes sociales. Estos sistemas brindan una solución generalmente fácil de usar, no demasiado costosa y escalable. Sin embargo, entre sus desventajas se encuentra que no cuentan con un único modelo de datos a nivel de sistema, en cuanto a la arquitectura se tiene poca estandarización de interfaces para servicios, tampoco se tiene una semántica estándar y esto trae consigo problemas de interoperabilidad (Jaramillo Valbuena & Londoño, 2015).

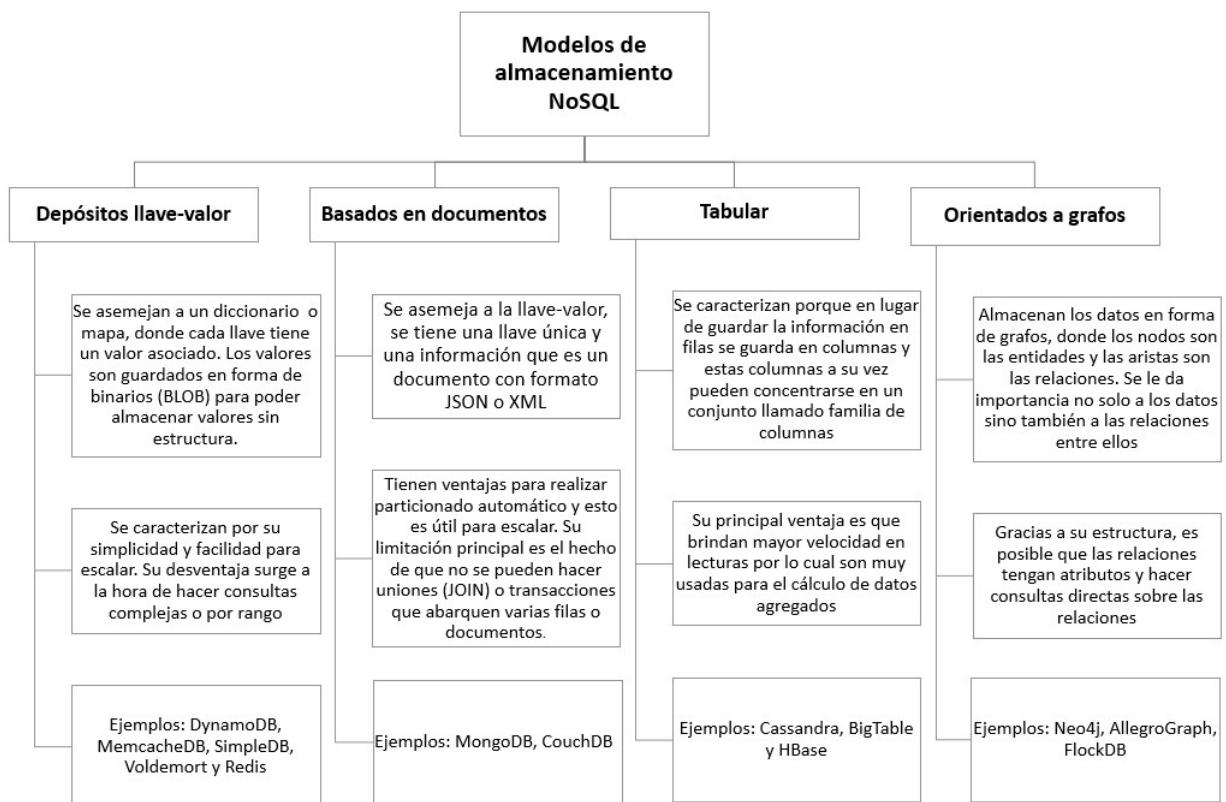


Figura 2-5: Modelos de almacenamiento NoSQL. Construido a partir de (Jaramillo Valbuena & Londoño, 2015)

2.3 Técnicas de análisis de datos

Existen diferentes tipos de técnicas de análisis datos, como las de tipo estadístico, gráfico, numérico – matemático, entre otras. En este aparte se describirán algunas técnicas que se basan en herramientas informáticas y en cierta medida en inteligencia artificial.

2.3.1 Consultas o análisis relacional

Las consultas relacionales hacen referencia a la extracción de conocimiento desde un almacenamiento de tipo de relacional. Estas consultas consideran modelos relacionales, en los cuales se tiene una estructura lógica con relaciones o tablas. Las relaciones tienen nombre y están formadas por atributos (columnas) de datos que presentan un tipo particular (enteros, flotantes, caracteres, etc). Para la generación de consultas relacionales se usa el lenguaje, preferiblemente, SQL. Las operaciones fundamentales del álgebra relacional son: selección, proyección, producto cartesiano, unión y diferencia de conjuntos. También se tienen la unión, intersección y operación de división, las cuales se expresan en términos de las siguientes operaciones básicas: count, sum, avg, min y max. Estas operaciones se conocen como operaciones de agrupamiento (Ramírez Romero, Patiño Ortiz, & Patiño Ortiz, 2015)

2.3.2 Procesamiento analítico en línea OLAP

Corresponde a las siglas en inglés de On-Line Analytical Processing, en español procesamiento analítico en línea. Este término fue introducido en el año 1993 por Codd, Codd y Salley con el apoyo de Arbor Software Corporation. OLAP comprende tres tipos de procesamiento de datos que se caracterizan entre otras cosas, por permitir el análisis multidimensional. “Dicho análisis consiste en modelar la información en medidas, dimensiones y hechos. Las medidas son los valores de un dato, en particular, las dimensiones son las descripciones de las características que definen dicho dato y los hechos corresponden a la existencia de valores específicos de una o más medidas para una combinación particular de dimensiones” (Abril Fraile & Pérez Castillo, 2007). Existen varios modos de almacenamiento OLAP, almacenamiento ROLAP (relational OLAP), almacenamiento MOLAP (multidimensional OLAP) y almacenamiento HOLAP (hibrid OLAP) (Tamayo & Moreno, 2006). En la Figura 2-6 se muestra un resumen con las principales características de los tipos de OLAP.

2.3.1 Minería de Datos

La minería de datos se puede definir como el proceso de extracción de conocimiento a partir de cúmulos de datos. Se suele utilizar el término minería de datos como sinónimo de descubrimiento de conocimiento, pero realmente no son sinónimos, la minería de datos es

solo un paso en el proceso de descubrimiento de conocimiento KDD (Knowledge Discovery in Databases) (Han et al., 2011). Los componentes de KDD se pueden apreciar en la Figura 2-7.

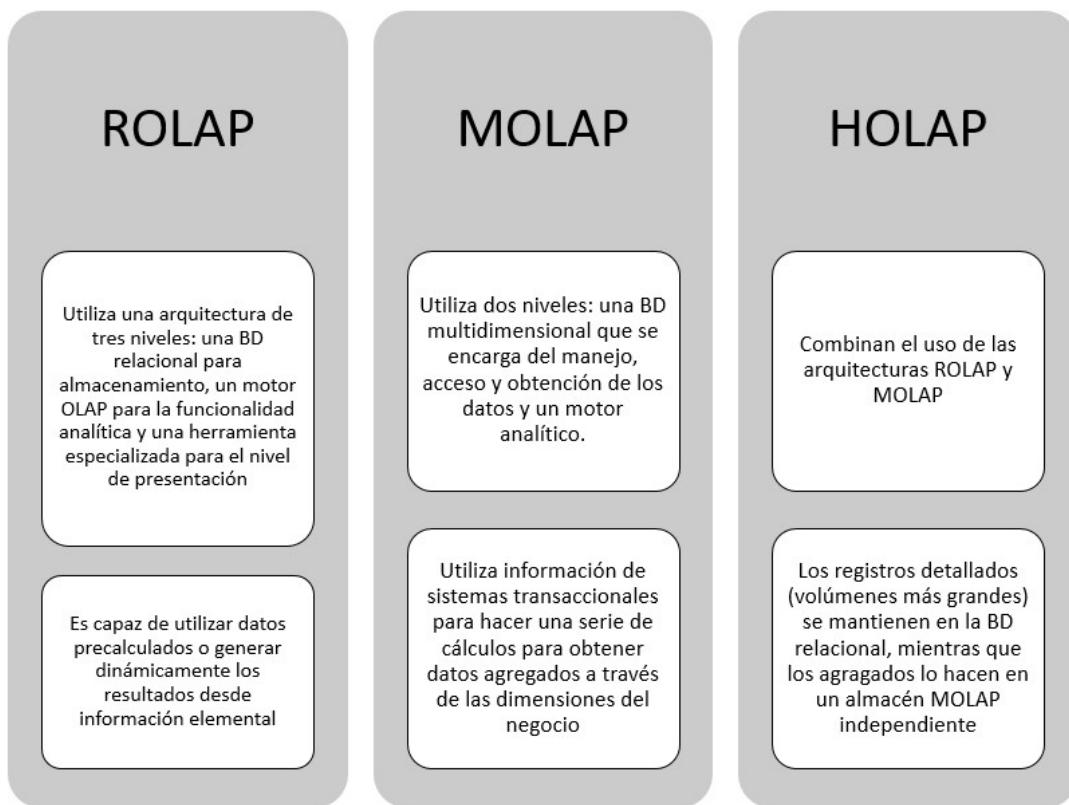


Figura 2-6: Tipos de OLAP. Elaboración propia

La minería de datos nace de la necesidad de conocer información útil a partir de los bases de datos o Datawarehouse, con el crecimiento de los datos disponibles, la inteligencia de negocios tuvo que dar paso a la aplicación de la minería de datos en soluciones empresariales y comerciales, puesto que de esta manera se permite el descubrimiento automático o semiautomático de información relevante a partir de estos cúmulos de datos. En las ciencias y la ingeniería existe un amplio rango de problemas y dominios de aplicación para la minería de datos (Grossman, Kamath, Kegelmeyer, Kumar, & Namburu, 2013). Se encuentran soluciones a partir de minería de datos para problemas de los campos de mercadeo, comercio, salud, predicción, transporte, meteorología, entre otros.

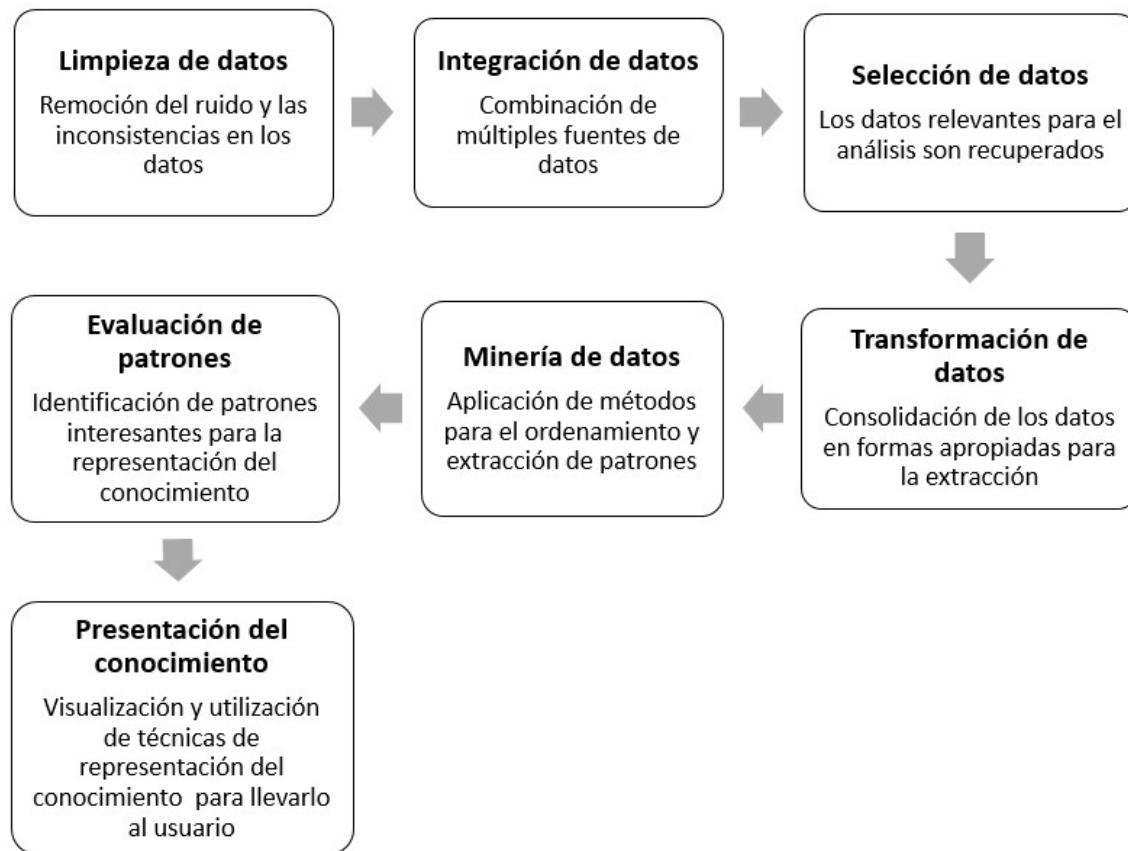


Figura 2-7: Esquema de los pasos del proceso de descubrimiento del conocimiento (KDD). Adaptado de (Han et al., 2011).

2.3.2 Machine Learning y Deep Learning

Machine Learning o aprendizaje máquina es un área de investigación bastante reconocida en las ciencias de la computación, principalmente comprende el descubrimiento de modelos, patrones y regularidades en los datos (Michalski, Bratko, & Bratko, 1998). El aprendizaje máquina puede ser visto desde dos enfoques, los simbólicos y los estadísticos. Los primeros trabajan aprendizaje inductivo de descripciones simbólicas, mientras que los segundos se centran en los métodos de reconocimiento de patrones o en la estadística. En los últimos años, el uso del aprendizaje máquina se ha extendido con rapidez (Domingos, 2012), se ven aplicaciones en dominios como detección de fraudes, sistemas de recomendación (Portugal, Alencar, & Cowan, 2015), detección de spam (Crawford, Khoshgoftaar, Prusa, Richter, & Al Najada, 2015), predicciones financieras (Lin, Hu, & Tsai,

2012), comercio y mercadeo (Dash & Dash, 2016), entre otros. Los algoritmos de aprendizaje máquina se clasifican en supervisados y no supervisados.

Por su parte, Deep Learning o aprendizaje profundo es un término general usado para referirse a una serie de arquitecturas de múltiples capas que son entrenadas usando algoritmos no supervisados. Una de las principales características es el modelado de abstracciones de alto nivel a través de métodos no supervisados, con el objetivo de conseguir una nueva representación de datos que contribuya a la tarea de predicción. Este enfoque ha sido aplicado con éxito en campos como la visión artificial, el reconocimiento de imágenes, el procesamiento del lenguaje natural y la bioinformática (Arel, Rose, & Karnowski, 2010). Deep Learning se muestra prometedor para el modelado de datos de series de tiempo a través de técnicas como las Máquinas Restringidas de Boltzmann o RBM (por sus siglas en inglés, Restricted Boltzmann Machine), RBM Condicionales, Autoencoders (autocodificadores), Redes Neuronales Recurrentes, Redes Neuronales Convolucionales, Modelos ocultos de Márkov o HMM (por sus siglas del inglés, Hidden Markov Model) (Längkvist, Karlsson, & Loutfi, 2014).

2.3.3 Tecnologías asociadas a *Big Data*

Para el análisis de datos masivos ha surgido todo un ecosistema que se fundamenta en el paradigma de “divide y vencerás”. Como principal solución se ha consolidado *Hadoop*.

2.3.3.1 *Hadoop*

Hadoop es una librería de Apache definida como un framework que permite hacer procesamiento de datos distribuido sobre volúmenes de datos de considerable tamaño sobre clúster. Está diseñado pensando en brindar poder de escalamiento desde un par de servidores hasta cientos de máquinas o nodos, las cuales manejan almacenamiento y procesamiento local (The Apache Software Foundation, 2016b). *Hadoop* ha tomado gran auge en los últimos años, convirtiéndose en una de las soluciones más empleadas a nivel organizacional para afrontar el tratamiento de *Big Data*. *Hadoop* cuenta con dos componentes esenciales, un sistema de archivos distribuido (HDFS) y *MapReduce*. Los pilares básicos de un HDFS se presentan en la Figura 2-8.

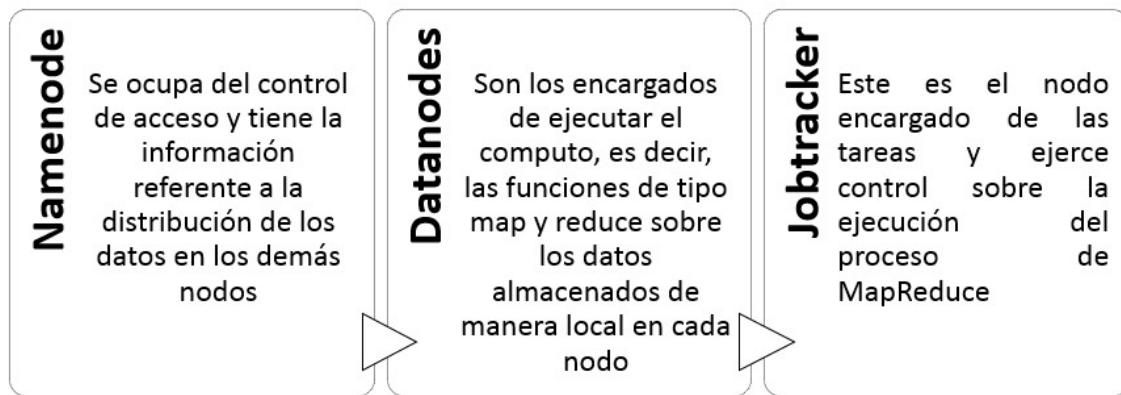


Figura 2-8: Pilares de un HDFS. Elaboración propia

Además, el HDFS cuenta con las siguientes características fundamentales:

- Tolerancia a fallos
- Acceso a datos en streaming
- Facilidad para el trabajo
- Modelo sencillo de coherencia
- Portabilidad de convivencia

Varios trabajos donde se han tomado *Hadoop* como base y se ha potencializado algunas de sus características o se ha fusionado con otra herramienta o tecnología. Ejemplos de esto se pueden encontrar en (Dittrich et al., 2010), (Bu, Howe, Balazinska, & Ernst, 2010) y (Prekopcsák, Makrai, Henk, & Gaspar-Papanek, 2011).

2.3.3.2 *MapReduce*

MapReduce es un modelo de programación que se ha asociado también a la implementación de estrategias de procesamiento de grandes conjuntos de datos que puede ser aplicado a una gran variedad de tareas del mundo real (Dean & Ghemawat, 2008). Este modelo de programación fue utilizado inicialmente por Google para resolver el problema de ranking de páginas (“Page Rank”). El modelo se basa en los siguientes conceptos: iteraciones sobre los datos de entrada, construcción de los pares clave-valor a partir de cada pieza de entrada, agrupación de los valores intermedios de acuerdo a las claves, iteración sobre los grupos resultantes y reducción de cada grupo (Lämmel, 2008).

En la Figura 2-9 se presenta el esquema de un proceso *MapReduce* y seguidamente, se hace una descripción de cada una de fases que involucra.

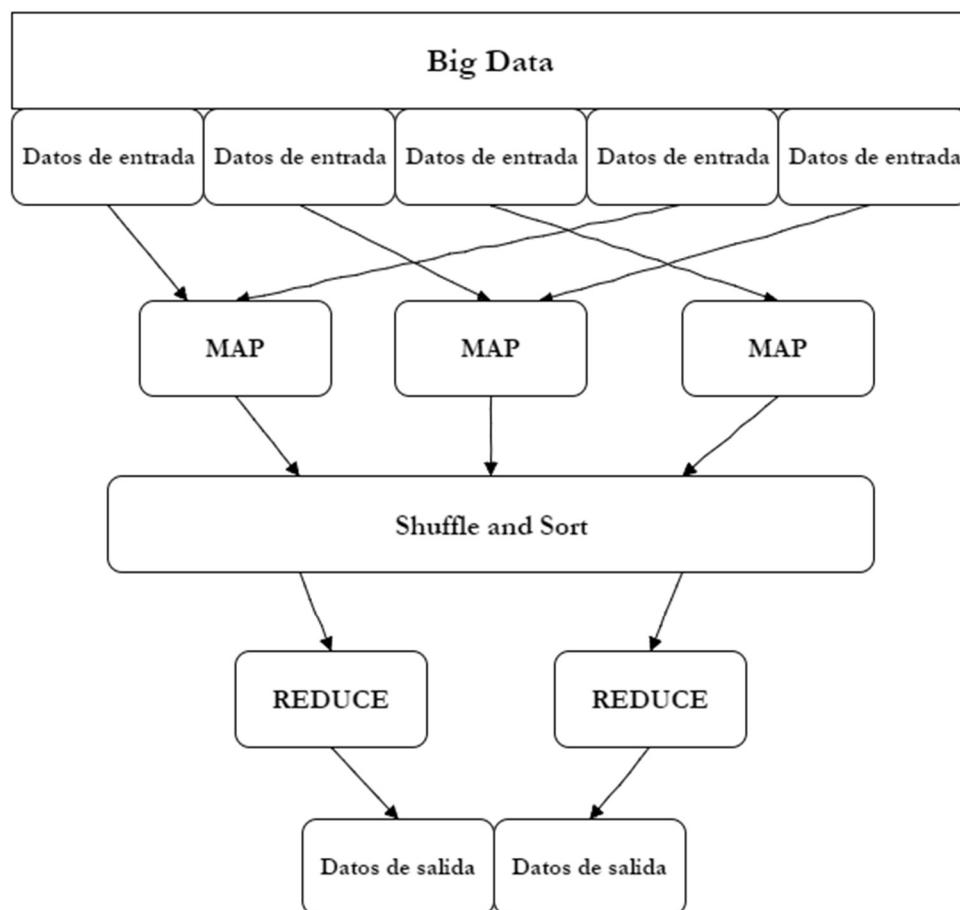


Figura 2-9: Esquema general de los procesos *MapReduce*. Elaboración propia

- **Map:** Se aplica en paralelo para cada uno de los ítems en la entrada de datos. Por medio de la tarea *Map* a cada llamada se asignará una lista de pares clave-valor (key-value). Por cada clave generada se crea un grupo, el framework agrupa todos los pares con la misma clave extraídos de todas las listas tratadas.
- **Reduce:** se aplica en paralelo para el grupo asociado a una clave. El resultado es la producción de una colección de valores para cada dominio.
- **Shuffle and sort:** tiene dos misiones, por una parte se encarga de ordenar por clave todos los resultados emitidos por los *mapper* y por otra parte recoge todos los valores intermedios pertenecientes a una clave para combinarlos en una lista asociada a ella.

Las características de *MapReduce* se resumen a continuación:

- Distribución y paralelización automáticas
- Tolerancia a fallos y a redundancias
- Transparencia
- Escalabilidad horizontal
- Localización de los datos
- Herramientas de monitorización

2.3.3.3 *Mahout*

Mahout es un proyecto de Apache que tiene como objetivo ofrecer un ambiente para la creación rápida de aplicaciones de aprendizaje máquina escalables y eficiente (The Apache Software Foundation, 2016a). *Mahout* ofrece una suite de algoritmos para clustering, categorización, filtrado colaborativo, clasificación y programación evolutiva; algunas de sus principales aplicaciones prácticas se enmarcan en la realización de clúster de documentos, recomendaciones y organización de contenidos (Ingersoll, 2009). El machine learning o aprendizaje máquina es el trasfondo principal de *Mahout* y corresponde a un subcampo de la inteligencia artificial que se centra en el mejoramiento de procesamientos computacionales a partir del análisis de experiencias previas. *Mahout* desde su aparición ha seguido siendo un proyecto en desarrollo, crecimiento y expansión; En (Ingersoll, 2012) se presenta una descripción de algunos de los más recientes algoritmos implementados en *Mahout*, resumiéndolos en la Tabla 3, la cual se muestra a continuación.

Tabla 2-1: Algunos algoritmos en *Mahout*. Tomado y adaptado de (Ingersoll, 2012)

Algoritmo	Descripción breve	Aplicaciones
Regresión logística, resuelta por gradiente estocástico descendiente (SGD)	Clasificador brillante, rápido, simple y secuencial, capaz de aprendizaje on-line en entornos exigentes	Recomendación de publicidad, clasificación de textos
Modelos ocultos de Markov (HMM)	Implementaciones secuenciales y paralelas del algoritmo clásico de clasificación diseñado para modelar procesos del mundo real cuando el proceso de generación subyacente es desconocido	Etiquetado de texto, reconocimiento del discurso

Descomposición de valor singular (SVD)	Diseñado para reducir el ruido en matrices grandes, haciendo con esto que sean más pequeñas y que sea más fácil trabajar con ellas	Clasificación para realizar selección de recursos automáticamente
Almacenamiento en clúster Dirichlet	Enfoque de almacenamiento en clúster basado en modelo, que determina la propiedad con base en si los datos se ajustan al modelo subyacente	Almacenamiento en clúster para datos con sobreposición o jerarquía
Almacenamiento en clúster espectral	Es una familia de enfoques similares que usa un enfoque basado en gráficas para determinar la membresía a clúster	Almacenamiento en clúster para conjuntos de datos grandes y no vistos
Almacenamiento en clúster Minhash	Utiliza una estrategia de hash para agrupar elementos similares, produciendo así clústeres	Clúster
Numerosas mejoras de recomendador	Co-ocurrencia distribuida, SVD, mínimos cuadrados alternantes	Recomendaciones en sitios de citas, e-commerce, recomendaciones de películas o de libros
Colocaciones	Implementación de colocación reducida por correlacionamiento	Encontrando frases estadísticamente interesantes en texto

2.4 Componente ambiental

Las dinámicas de los recursos naturales y el medio ambiente generan una cantidad de datos considerable, los cuales pueden ser medidos, almacenados y analizados. Tratar estos datos exige de un gran esfuerzo, pero pueden contribuir de manera significativa a la generación de conocimiento, de patrones de comportamiento, de alertas de prevención y de políticas ambientales para apoyo a la toma de decisiones.

Dentro de las variables del ambiente que se suelen monitorear y que generan datos de forma continua se encuentran las hidrometeorológicas. La meteorología es la ciencia que está encargada del estudio de la atmósfera, incluyendo las propiedades de esta y los fenómenos que ocurren en ella. Este estudio de la atmósfera se da por medio del conocimiento de una serie de variables como temperatura del aire, presión atmosférica, humedad, precipitación, entre otras, las cuales cambian de acuerdo al lugar y tiempo en que se midan. La importancia de los estudios meteorológicos radica en la influencia que tienen las condiciones atmosféricas, dadas por el comportamiento de variables como las mencionadas, que se presentan en un momento y lugar específico, lo cual es un

condicionante para diferentes actividades como agricultura, el transporte de mercancías, la minería, entre otras. (Rodríguez Jiménez, Capa, & Portela Lozano, 2004).

2.4.1 Redes y estaciones de monitoreo hidroclimatológico

Cuando se dio inicio a las mediciones meteorológicas, las observaciones eran físicas, no se contaba con instrumentos, lo cual hacia que estas mediciones fuesen poco precisas. Sin embargo cuando surgió la disponibilidad de instrumentación, la meteorología se volvió una ciencia experimental formal (Orozco-Alzate, Velez-Upegui, & Duque-Mendez, 2014). Ahora bien, en la actualidad se cuenta con una variedad de instrumentos y sensores bastante sofisticados que permiten cada vez hacer mediciones con mayor precisión y abarcar las variables necesarias. Estos instrumentos son emplazados en estaciones de monitoreo y las estaciones a su vez hacen parte redes de monitoreo.

Se pueden realizar diferentes clasificaciones de las estaciones de monitoreo meteorológico e hidrometeorológico, dependiendo por ejemplo, del tipo de mediciones que toman. Una posible clasificación es: estaciones principales y estaciones ordinarias. Una estación principal es aquella donde se hacen observaciones de lluvia, temperatura del aire, temperaturas máximas y mínimas, humedad, viento, radiación, brillo solar, evaporación, cantidad de nubes y fenómenos especiales. Una estación ordinaria es aquella donde se hacen observaciones de temperatura del aire y precipitación primordialmente. Una estación hidrometeorológica es aquella que realiza mediciones de nivel y caudal, ya que se ubica sobre una fuente hídrica y también reporta datos de precipitación y temperatura del aire. Por su parte, una estación hidrométrica, es aquella que mide nivel y caudal.

Una red hidrometeorológica hace referencia a un conjunto de estaciones, que transmiten datos de algunas de las variables hidroclimatológicas mencionadas hacia una estación central en la cual son almacenados y posteriormente procesados y analizados por expertos en el tema.

2.4.2 Algunas variables hidrometeorológicas

Las variables relacionadas con el componente atmosférico caracterizan el clima de una región específica, algunas de las variables predominantes son temperatura del aire, precipitación, presión atmosférica, dirección y velocidad del viento (Ocampo López & Vélez Upegui, 2015). A continuación se describirán algunas de las variables hidrometeorológicas medidas por sensores ubicados en las estaciones de monitoreo.

- Temperatura del aire: es una de las principales magnitudes utilizadas para describir el estado de la atmósfera. Dicha magnitud está determinada por la rapidez del movimiento de las partículas que constituyen la materia, a mayor agitación en las partículas mayor temperatura (Rodríguez Jiménez et al., 2004). Existen diferentes escalas para expresar la temperatura, una de las más usadas son los grados centígrados, y el instrumento utilizado para medir esta magnitud es el termómetro.
- Precipitación: se define como los productos líquidos o sólidos de la condensación de vapor de agua que caen de las nubes o son depositados desde el aire sobre la tierra. La cantidad total de precipitación que llega al suelo en un periodo determinado se expresa en términos de la profundidad vertical de agua la cual cubre una proyección horizontal de la superficie de la Tierra (Organización Mundial de Meteorología OMM, 2008). La unidad en la que se expresa este indicador son milímetros (mm).
- Brillo solar: es también conocido como insolación y es medido por medio de un Heliógrafo (Ocampo López & Vélez Upegui, 2015), representa el tiempo durante el cual incide luz solar directa sobre alguna localidad, entre el alba y el atardecer. El total de horas de brillo solar de un lugar es uno de los factores que determinan el clima de esa localidad (ETESA, 2009).
- Dirección viento: está definida por el punto del horizonte del observador desde el cual el viento sopla. En la actualidad, se usa internacionalmente la rosa de vientos dividida en 360°. («Sistema de información Ambiental de Colombia - SIAC», 2011).

- Radiación solar: es la energía emitida por el sol, que se propaga en todas las direcciones a través del espacio mediante ondas electromagnéticas. Esta energía es el motor que determina la dinámica de los procesos atmosféricos y el clima («Sistema de información Ambiental de Colombia - SIAC», 2011).
- Presión barométrica: es la fuerza por unidad de superficie que ejerce el aire sobre todos los cuerpos debido a la acción de gravedad. Esta magnitud está influenciada por otras variables como por ejemplo la altitud. La humedad, la temperatura, la situación geográfica, entre otras. La unidad de medida de esta magnitud es el Pascal. (Rodríguez Jiménez et al., 2004).
- Humedad relativa: se define como la cantidad de vapor de agua que está en el aire. Dicha cantidad es variable y depende de factores como la lluvia, la cercanía al mar, la vegetación, etc. Existen diferentes formas de expresar la humedad, sin embargo, la más usada es la humedad relativa y se expresa en porcentaje (%), por lo que va de 0 a 100%, cuando alcanza este último valor significa que la masa de aire ya no puede almacenar más vapor de agua, por lo cual la cantidad de vapor adicional se empieza a convertir en agua o en cristales de hielo. (Rodríguez Jiménez et al., 2004).

Capítulo 2

3. Estado del arte

En este capítulo se presenta la revisión de algunos trabajos relacionados con el área que enmarca el problema de investigación. Estos trabajos han sido analizados desde sus fortalezas, limitaciones y en algunos casos oportunidades. Así mismo, se ha realizado una clasificación de los trabajos bajo tres grupos, en un primer grupo se tienen en cuenta los trabajos relacionados con tendencias en *Big Data*; en el segundo grupo se hace un compendio de los trabajos relacionados con el diseño de modelos de almacenamiento y análisis de datos y el tercer grupo apunta a trabajos relacionados con la aplicación de técnicas de inteligencia artificial para la realización de pronósticos de variables ambientales, principalmente de precipitación. A continuación se presentarán los tres grupos de trabajos y una síntesis de cada uno.

3.1 Primer grupo: tendencias en *Big Data*

Este primer grupo comprende trabajos que presentan una revisión del estado actual en temas relacionados con *Big Data*, las tendencias y enfoques en el desarrollo de investigaciones en este campo.

- (Wu, Zhu, Wu, & Ding, 2014) presentan la importancia y los retos que actualmente se vienen dando con el consumo y creación de información a través de la red. En este sentido las aplicaciones de *Big Data*, son las que están presentando mayor capacidad de capturar, manejar y procesar estos grandes volúmenes de datos. Los autores proponen en su trabajo a “HACE”, un teorema con el cual buscan describir las características de la revolución de *Big Data*; así mismo, plantean los componentes que debe considerar un marco de trabajo para el procesamiento de *Big Data* desde la perspectiva de la minería de datos, el cual debe estar orientado a los datos, considerar la agregación de fuentes de información, la minería y el

análisis, el modelado de interés de los usuarios, y aspectos de seguridad y privacidad.

El teorema “HACE” se refiere a que *Big Data* comienza con un gran volumen de datos **heterogéneos** y provenientes de fuentes **autónomas** con control distribuido y descentralizado, y que trata de explorar relaciones **complejas** y **cambiantes** entre los datos, HACE corresponde a las iniciales de Heterogeneous, Autonomous, Complex y Evolving, siendo estas características, las que se convierten en un gran desafío para descubrir conocimiento útil desde el *Big Data*. La heterogeneidad se refiere a los diferentes tipos de representaciones para los mismos individuos, y la diversidad de características se refiere a la variedad de las características relacionadas a la hora de representar cada observación particular. Las fuentes de datos autónomas con control distribuido y descentralizado son, según los autores, la principal característica de las aplicaciones de *Big Data*; al ser autónomas, cada fuente de datos tiene la capacidad de generar y recopilar información sin la participación de un ente de control centralizado. Esto es similar a la World Wide Web (WWW) donde cada servidor web proporciona una cierta cantidad de información y cada servidor es capaz de funcionar plenamente sin necesidad de depender de otros servidores. Finalmente, con el incremento del volumen de datos, también se da la complejidad y el cambio constante en las relaciones entre los datos, un claro ejemplo de esta característica está en las redes sociales, como Facebook o Twitter, donde las correlaciones entre individuos son complicadas de representar.

Un marco de trabajo para el procesamiento de *Big Data* presenta ciertos desafíos de investigación, los cuales se pueden reunir en una estructura de tres niveles, como se aprecia en la Figura 3-1, en la parte central, la "plataforma de minería de *Big Data*" (nivel I), que se enfoca en el acceso a los datos de bajo nivel y computación. Los desafíos en el intercambio de información y la privacidad, los dominios de aplicación de *Big Data* y el conocimiento forman el nivel II, que se concentra en la semántica de alto nivel, las aplicaciones de dominio de conocimiento y los problemas de privacidad del usuario. Por su parte en el nivel III (el círculo exterior) se presentan los desafíos en los actuales algoritmos de minería.

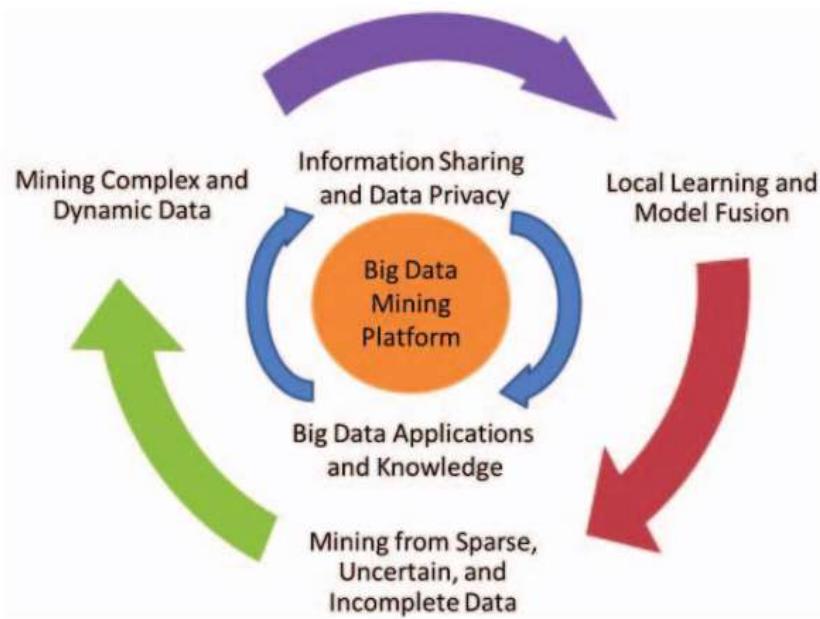


Figura 3-1: Marco de trabajo para el procesamiento de *Big Data*. Tomado de (Wu et al., 2014)

Fortalezas

- Se describe la importancia que tiene el análisis de los datos generados a través de la red y la complejidad que han ido adquiriendo los mismos.
- Se establecen claramente los desafíos y aspectos primordiales que deben ser tratados a la hora de formular un modelo para el tratamiento de *Big Data* enfocado en la minería de datos.

Oportunidades

- Se plantea la necesidad de diseñar un sistema o modelo que permita vincular datos no estructurados a través de sus relaciones complejas y con esto llegar a establecer patrones útiles.
- El crecimiento en el volumen de datos y en las relaciones entre elementos debe ayudar a legitimar los patrones para predecir tendencias.
- (Jaramillo Valbuena & Londoño, 2015) presentan una revisión del estado del arte en cuanto a sistemas de almacenamiento para grandes volúmenes de datos, incluyendo un comparativo entre los sistemas de administración de bases de datos (DBMS) tradicionales y los nuevos enfoques NoSQL (Not Only SQL). En el trabajo se considera la necesidad de que estos sistemas sigan garantizando

características como: escalabilidad, fiabilidad, durabilidad, tiempos de respuesta, interfaces de consulta, esquemas de particionamiento y estructura o carencia de esta. En el trabajo se describen los cuatro modelos de almacenamiento NoSQL, depósitos llave-valor, basado en documentos, tabular y orientados a grafos. Los autores afirman que los sistemas NoSQL se adecuan a casos en los que se necesita atender a muchos usuarios sin perder rendimiento, como puede pasar en el caso de las redes sociales. Por su parte, recomiendan los sistemas de bases de datos relacionales cuando se trata de garantizar integridad referencial, se requiere el uso de conexiones entre servidores y clientes, consultas arbitrarias, estandarización, herramientas de análisis y pruebas de rendimiento. En Tabla 3-1 se presenta un análisis de algunas de las principales implementaciones de sistemas de almacenamiento.

Tabla 3-1: Comparativo de implementaciones de sistemas de almacenamiento. Tomado de (Jaramillo Valbuena & Londoño, 2015)

Base de datos		Consistencia (+Strong (/)BASE (-Tunable))	Modelo de datos	Arquitectura	Particionamiento
Relacionales	Vertica	+	Tabular	Share-nothing	Mixta-- Híbrido Fila/Columna
	Teradata	+	Tabular	Share-nothing	Mixta-- Híbrido Fila/Columna
	HadoopDB	+	Tabular	Share-nothing	*
NoSQL--Tabular	Cassandra	+/-	Columnas, grupo de columnas correspondientes a una clave (supercolumnas)	Share-nothing	Hashing consistente
	BigTable	+	Mapa multidimensional, cada valor del mapa está indexado por 3 valores: <i>row key, column key y timestamp</i>	Share-nothing	Rango
	HBase	+	Grupo de columnas (clon de BigTable)	Share-nothing	Rango
NoSQL--Documentos	MongoDB	+/-	Documentos con información semiestructurada almacenada en colecciones	*	Rango
	CouchDB	/	Documentos como una lista de ítems llamados JSON	*	Hashing consistente
NoSQL Clave-valor	Dynamo	+/-	Grupo de parejas Clave-Valor	Share-nothing	Hashing consistente
NoSQL Grafos	Neo4j	+	Grafo con nodo y aristas	*	*
	AllegroGraph	+	Grafo con nodo y aristas	*	*

Fortalezas y oportunidades

- Se resaltan las ventajas que presentan tanto los DBMS tradicionales como los enfoques NoSQL, mostrando que existe incluso la posibilidad de pensar en la construcción de sistemas híbridos, que contengan varios almacenes de datos atendiendo a las necesidades de cada set de datos.
- Se muestran varios temas latentes de investigación, que incluyen la necesidad de atender falencias en los sistemas NoSQL como por ejemplo, los problemas que presentan los almacenes clave-valor para realizar consultas complejas, bases de datos orientadas a grafos que no pueden realizar particionamiento, la incapacidad de realizar uniones o transacciones que abarquen varias filas o documentos por parte de bases de datos documentales o tipo BigTable.
- Por su parte, en (Jin, Wah, Cheng, & Wang, 2015) se describe como en el área de la industria y los negocios se ha presentado una explosión en el número de datos, causada principalmente por el rápido desarrollo del internet, nuevos conceptos como el internet de las cosas y la computación en la nube. *Big Data* se ha constituido como un “tópico caliente” que atrae la atención no solo de la industria, sino también de la academia y del gobierno. Los autores presentan desde diferentes perspectivas el significado y las oportunidades que nos brinda el ecosistema *Big Data* y dan una serie de condiciones necesarias para que un proyecto de *Big Data* sea exitoso. En primer lugar, se deben tener claros los requerimientos independientemente de si son técnicos, sociales o económicos. En segundo lugar, para trabajar de forma eficiente con *Big Data* se requiere explorar y encontrar la estructura central o el kernel de los datos a ser procesados, ya que al tener esto se puede caracterizar el comportamiento y las propiedades subyacentes a *Big Data*. En tercer lugar se debe adoptar un modelo de administración top-down, se puede considerar también un modelo bottom-up, sin embargo solo serviría cuando se trata de problemas específicos, y luego tratar de unirlos para formar una solución completa es complejo. Por último, los autores exponen la necesidad de abordar desde los proyectos *Big Data* soluciones integradas, no con esfuerzos aislados.

Fortalezas y oportunidades

- Se hace una breve revisión de las oportunidades e importancia de *Big Data*, pero se enfatiza en cómo hacer un proyecto de *Big Data* exitoso. Para ello, se da una serie de recomendaciones como tener claridad en los requerimientos, encontrar el centro de los datos a procesar, caracterizar el comportamiento y propiedades del problema, ya que cada dominio de datos es específico.
- (Sagiroglu & Sinanc, 2013) presentan la revisión de varios aspectos relacionados con *Big Data*, tales como contenido, alcance, métodos, ventajas, desafíos, ejemplos y privacidad de los datos. La revisión realizada por los autores muestra que incluso con las herramientas y técnicas disponibles en la actualidad y la literatura al respecto, existen muchos puntos a ser considerados, desarrollados, mejorados y analizados. Es claro que la cantidad de datos ha ido en aumento, lo cual exige que también las técnicas de análisis y tratamiento de datos se hagan más competitivas, el reto no es sólo para recoger y gestionar el gran volumen y diferentes tipos de datos, sino también para extraer valor significativo de estos. Se presentan como las principales barreras para la implementación de analíticas de *Big Data*: la carencia de expertos en el tema de *Big Data*, el costo, el manejo de la privacidad en la manipulación de los datos, la dificultad en el diseño de sistemas de análisis, la falta de software que soporte grandes bases de datos permitiendo análisis con tiempos de procesamiento rápido, los problemas de escalabilidad, la incapacidad de hacer que *Big Data* sea utilizable por usuarios finales, la falta de rapidez en la carga de datos con los sistemas de gestión de bases de datos actuales y la ausencia de un modelo de negocio convincente y rentable en torno al tema. En el documento se hace también un comparativo entre dos de los principales marcos de trabajo desarrollados para el tratamiento de *Big Data*. *Hadoop* es una solución basada en archivos distribuidos, de código abierto, que se utiliza para almacenar, procesar y analizar grandes volúmenes de datos. *Hadoop* cuenta con dos componentes principales, un HDFS (sistema de archivos distribuidos *Hadoop*) y *MapReduce*, una poderosa técnica de programación en paralelo para procesamiento distribuido sobre clústeres. El segundo marco de trabajo revisado es HPCC (High Performance Computing Cluster), es una plataforma de código abierto para el manejo intensivo de datos en sistemas

distribuidos, que proporciona servicios de gestión de flujos de trabajo sobre *Big Data*. Una de las diferencias con *Hadoop*, es que en HPCC el modelo de datos es definido por el usuario. HPCC cuenta con tres componentes principales: Thor, un motor de ETL paralelo y masivo que permite la integración de datos a gran escala; Roxie, un motor de entrega de datos que permite de manera eficiente la recuperación de datos y consultas estructuradas desde múltiples usuarios; el lenguaje de control ECL, el cual distribuye automáticamente la carga de trabajo entre nodos, cuenta con una librería de aprendizaje automático y proporciona un lenguaje de programación optimizado para operaciones y consultas sobre *Big Data*. La Figura 3-2 muestra una comparación entre las arquitecturas de estos dos marcos de trabajo.

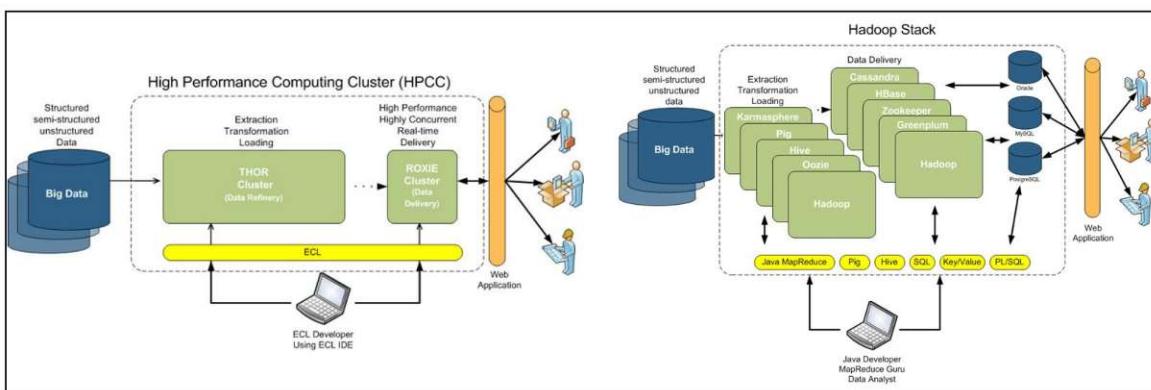


Figura 3-2: Comparación entre HPCC y *Hadoop*. Tomado de (Sagiroglu & Sinanc, 2013)

Fortalezas

- Los dos frameworks revisados por los autores presentan ventajas que incluyen la agilidad para el procesamiento de información, la optimización de recursos y la flexibilidad de uso; sin embargo, traen consigo altos costos en habilidades y requieren de un mayor tiempo de adaptación.
 - La revisión realizada por los autores deja varios temas claros y a disposición de los investigadores para su análisis y aborde posterior.

Limitaciones

- Mientras se da una sustitución en las plataformas de análisis, puede que los grandes volúmenes de datos no se logren contener, no se pueda apoyar a los modelos analíticos necesarios, la carga de datos sea demasiado lenta o los requisitos en cuanto a plataformas de análisis no puedan ser cubiertos.
- (Chen, Mao, & Liu, 2014) revisan los antecedentes y estado del arte de *Big Data*. Los autores analizan algunas tecnologías relacionadas con *Big Data* como computación en la nube, internet de las cosas, centros de datos y *Hadoop*. También se enfocan en la discusión de los desafíos técnicos y adelantos en cada una de las fases de *Big Data*: generación, adquisición, almacenamiento y análisis de datos. El análisis de *Big Data* tiene que afrontar muchos desafíos, se requieren considerables esfuerzos investigativos, los cuales se pueden agrupar en los problemas abiertos mostrados en la Figura 3-3.



Figura 3-3: Problemas abiertos en *Big Data*. Construido a partir de (Chen et al., 2014)

Así mismo, los autores presentan algunas perspectivas, no solo en cuanto al impacto social y económico de *Big Data*, sino que influyen en cada uno de los aspectos de la vida que se puedan imaginar. Ellos afirman que sin tratar de predecir el futuro, el panorama de *Big Data* se concentrará en:

- ✓ Datos con escalas y diversidad cada vez mayores y estructuras mucho más complejas
- ✓ El rendimiento de los recursos de datos
- ✓ *Big Data* promoverá la fusión transversal de la ciencia
- ✓ Retos de visualización de datos
- ✓ Orientación a los datos
- ✓ *Big Data* desencadenará una revolución del pensamiento

Fortalezas

- La revisión presentada por los autores deja claro que es necesario que se siga dando investigación aplicada en torno a las problemáticas de *Big Data*.
- Es bastante completo este trabajo, por lo cual además de presentar un estado del arte en el tema, sirve como marco teórico referente.

Limitaciones

- Aún quedan muchos vacíos conceptuales y teóricos que dejan ver que *Big Data* suele ser tomado más como una moda que como un enfoque científico.
- (O'Leary, 2013) hace énfasis en la utilización que se le ha dado a la Inteligencia Artificial (IA) para facilitar la captura y estructuración de grandes volúmenes de datos y también como se ha implementado para el análisis de estos. Se presentan algunas preocupaciones respecto a la integración de la IA con *Big Data*, que no se resuelven solo con pensar en la distribución y paralelización, sino que requieren otros análisis. La IA para el tratamiento de *Big Data* permite la delegación de tareas complejas de reconocimiento de patrones, aprendizaje y otras tareas basadas en enfoques computacionales, la IA contribuye a la velocidad en la manipulación de los datos, facilitando la toma de decisiones rápidas. Por ejemplo, muchas operaciones de la bolsa son hechas por sistemas basados en IA en lugar de personas, la velocidad de las operaciones puede aumentar y una transacción puede conducir a otras. Existen varios problemas emergentes asociados a la IA y *Big Data*, en primer lugar, la naturaleza de algunos de los algoritmos de machine-

learning son difícilmente usados en ambientes como *MapReduce*, por lo cual se requiere de su adaptación. En segundo lugar, *Big Data* trae consigo datos “sucios”, con errores potenciales, incompletos o de diferente precisión; la IA puede ser usada para identificar y limpiar estos datos sucios. En tercer lugar, la visualización de los datos; con la IA se puede lograr incluir la captura de capacidades de visualización de conocimiento para facilitar el análisis de datos; otro enfoque es hacer aplicaciones inteligentes de visualización para determinados tipos de datos. En cuarto lugar, ya que las tecnologías de almacenamiento evolucionan, es cada vez más factible proporcionar a los usuarios, casi en tiempo real, análisis de bases de datos más grandes, lo que acelera las capacidades de toma de decisiones.

Fortalezas

- Investigadores en el campo de la IA se han interesado en la creación de aplicaciones que permitan analizar datos no estructurados y que los resultados puedan ser categorizados o estructurados para interactuar con otras aplicaciones.
- Las empresas están empezando a utilizar *Big Data*, ya que con esto crean valor porque en esto se da solución a nuevos problemas y a problemas existentes pero de forma más rápida o económica.
- La IA puede ayudar a crear valor al proporcionar a las empresas un análisis inteligente de los datos e interpretaciones a la cada vez mayor cantidad de datos no estructurados disponibles.
- (Gandomi & Haider, 2015) presentan una descripción consolidada del concepto de *Big Data*, partiendo de las definiciones dadas por profesionales y académicos del campo como se puede ver en la Figura 3-4. Sin embargo, el artículo se concentra en revisar los métodos de análisis usados para *Big Data*. Se destaca que *Big Data* no tiene un verdadero sentido si solo se trata de un gran cúmulo de datos, su valor potencial se desbloquea sólo cuando estos datos son aprovechados para impulsar la toma de decisiones. Para ello es necesario mover y dar significado al *Big Data*, esto se puede hacer por medio de dos subprocessos principales: la gestión y análisis de datos; la gestión de datos implica procesos y tecnologías de apoyo para adquirir, almacenar, preparar y recuperar los datos para su análisis. El análisis, por su parte,

se refiere a las técnicas utilizadas para analizar y adquirir inteligencia a partir de *Big Data*. Los métodos de análisis de *Big Data* a los que hacen referencia los autores se enfocan en los tipos de datos tratados, por lo que se describen analíticas de texto, analíticas de audio, analíticas de social media y analíticas predictivas. Estas últimas, las predictivas, se basan principalmente en los métodos estadísticos; sin embargo, hay algunos factores que requieren el desarrollo de nuevos métodos estadísticos para *Big Data*. En primer lugar, los métodos estadísticos convencionales se concentran en una pequeña muestra de la población y los resultados se generalizan a toda la población; pero para el caso de *Big Data*, las muestras son enormes y representan la mayoría o la totalidad de la población. En segundo lugar, en términos de eficiencia de cómputo muchos métodos convencionales para muestras pequeñas no se logran escalar hasta *Big Data*. El tercer factor corresponde a algunos de los rasgos distintivos de *Big Data*: la heterogeneidad, la acumulación de ruido, las falsas correlaciones y la endogeneidad incidental.

Fortalezas

- En el documento se refuerza la necesidad de elaborar nuevas herramientas para el análisis predictivo sobre *Big Data*.
- Los avances tecnológicos en temas de almacenamiento y computación han permitido que se pueda tener una captura rentable de *Big Data* y de su valor informativo, todo de forma oportuna.
- Se observa una proliferación en la adopción de analíticas que antes de la era de *Big Data* no eran económicamente viables.

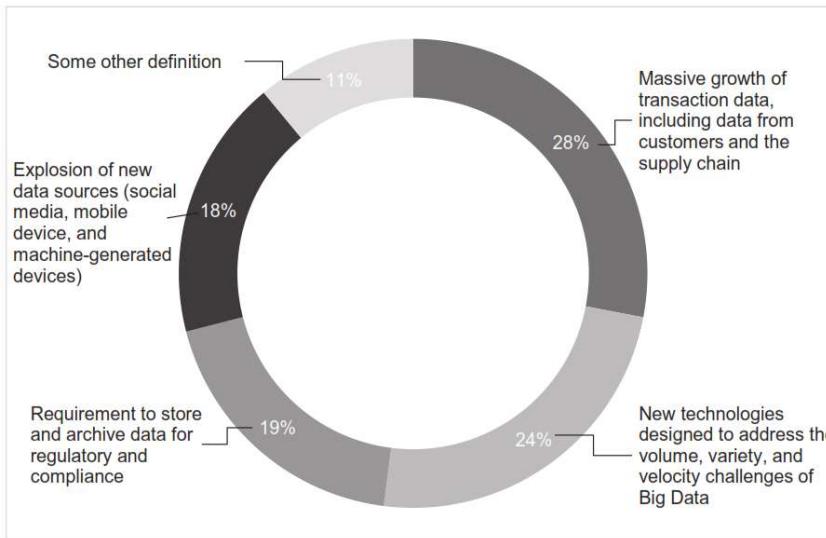


Figura 3-4: Definiciones de *Big Data* basadas en una encuesta en línea realizada a 154 ejecutivos globales en abril de 2012. Tomado de (Gandomi & Haider, 2015)

- (Chen & Zhang, 2014) presentan *Big Data* y sus aplicaciones, las oportunidades y desafíos de estas tecnologías, así como también técnicas de última generación que se han adoptado para hacer frente a los problemas de *Big Data*, estas se pueden apreciar en la Figura 3-5. Se discuten algunas metodologías utilizadas para tratar cantidades considerables de datos como son la computación granular, la computación en la nube, la computación bio-inspirada y la computación cuántica. Destacan el papel que han jugado los datos como impulsadores de diferentes campos científicos como la astronomía, la meteorología, la bioinformática y la biología computacional. Dichos campos basan gran parte de su descubrimiento científico en el análisis de grandes volúmenes de datos. Otro de los aportes significativos, es la descripción de los principios para el diseño de sistemas *Big Data*. Estos son: 1. Buenas arquitecturas y frameworks (son necesarios y de alta prioridad); 2. Soporte a una variedad de métodos analíticos; 3. No limitar a un tamaño definido para todo; 4. Conducción del análisis de los datos; 5. Procesamiento distribuido; 6. Almacenamiento de los datos distribuido; 7. Coordinación entre las unidades de procesamiento y de datos.

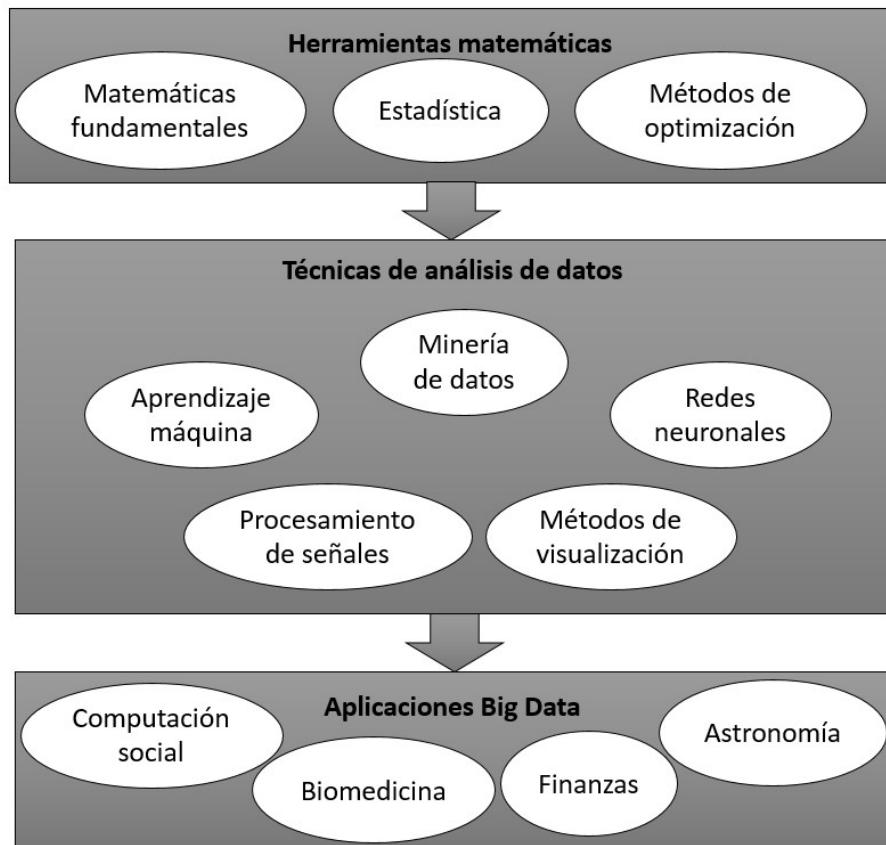


Figura 3-5: Técnicas *Big Data*. Adaptado de (C. L. P. Chen & Zhang, 2014)

Fortalezas

- Muestra *Big Data* como una nueva era que implica un reto para la innovación, competitividad, productividad y un nuevo movimiento en el campo científico.
- Muestra que el análisis de *Big Data* está todavía en una etapa inicial, a pesar de que se cuenta con técnicas y herramientas, estas aún son limitadas, complejas en su uso o costosas para ser aplicadas en la resolución de problemas reales.

3.2 Síntesis de los trabajos del primer grupo

En este primer grupo se revisaron algunos trabajos que permitieron determinar el estado actual del enfoque *Big Data* y de las tendencias que giran en torno a este, incluyendo el

planteamiento de varios campos de investigación que se encuentran abiertos, principalmente relacionados con la optimización de los sistemas de almacenamiento para grandes volúmenes de datos, los cuales todavía presentan falencias en cuanto al tratamiento de distintos tipos de datos a la vez, la optimización de consultas complejas y operaciones sobre los datos. También se ve la diversidad de planteamientos que presentan los autores en cuanto al concepto de *Big Data* y las características que este debe atender, es claro que el tema ha tomado un carácter de moda mundial y que se ha dejado de asociar solo a la característica de gran tamaño. Se ven también posibilidades de explorar la aplicación de *Big Data* a nuevos dominios de datos, ya que actualmente se han concentrado en social media, medicina, bioinformática y seguridad principalmente.

3.3 Segundo grupo: almacenamiento y análisis de datos

En el segundo grupo se han recogido algunos trabajos que presentan el diseño y/o implementación de modelos para el almacenamiento de grandes volúmenes de datos y procesos de análisis sobre estos.

- (Duque-Méndez et al., 2014) presentan el desarrollo de un modelo conceptual para la creación de una bodega de datos hidroclimatológicos, siguiendo un esquema en estrella y con el fin de aplicar análisis OLAP (On-Line Analytical Processing) para determinar algunos patrones y descubrir relaciones entre las variables. El modelo es probado con datos recolectados por las estaciones de monitoreo de la ciudad de Manizales y del departamento de Caldas, que cuentan con datos históricos, algunas incluso de 50 años atrás. A partir de este modelo se extraen algunos análisis los cuales pueden ser revisados a través de gráficas y que incluyen el comportamiento y relación entre las diferentes variables para períodos de tiempo diversos como días, años, lustros, entre otros. Los autores resaltan la importancia de realizar un proceso de ETL previo al tratamiento y análisis de los datos, con el fin de contar con una fuente de datos confiable.

Fortalezas

- El modelo conceptual para llegar al almacenamiento de los datos en la Bodega de Datos, incluyó una estrategia para la limpieza y unificación de los datos, lo que garantiza la confiabilidad de los mismos antes de realizar los procesos de extracción de conocimiento.
- Las fuentes de datos estudiadas presentan un buen número de datos, lo anterior hace que se puedan encontrar resultados más sólidos.

Limitaciones

- La alimentación de la bodega de datos se hizo con procesos manuales, lo cual hace que se dificulte mantenerla actualizada debido al gran número de datos que transmiten las estaciones y al futuro crecimiento de las redes de monitoreo.

Oportunidades

- El modelo de almacenamiento a través de la bodega de datos en estrella permite incluir en un futuro nuevas dimensiones y variables en la tabla central, lo cual puede permitir que se amplíen los resultados obtenidos de los análisis y que se encuentren nuevos patrones y conocimiento.
- Data Minind Meteo (DMM) es un proyecto presentado por (Bartok, Habala, Bednar, Gazak, & Hluchý, 2010) con el cual se proyecta hacer una contribución a la investigación de modelos y métodos para la detección y predicción de fenómenos meteorológicos, esto por medio de la creación de un modelo que permita el tratamiento de datos ambientales provenientes de mediciones relacionadas con la meteorología (principalmente niebla y nubosidad baja), incluyendo la medición y recolección de datos en tiempo real, el procesamiento y almacenamiento de datos y finalmente el análisis para dar soporte a la toma de decisiones. La propuesta para la integración de los datos y los procesos de minería se aprecia en la Figura 3-6. Se presentan los métodos y tecnologías que piensan usar los autores para hacer la integración de los datos de entrada, los cuales están distribuidos en servidores de diferentes proveedores. Los métodos de detección y predicción meteorológica se basan en técnicas estadísticas y climatológicas combinadas con el descubrimiento de conocimiento y la minería de datos. Los datos con que se

alimentará el modelo incluyen: imágenes de radar meteorológico, datos de estaciones meteorológicas "en bruto", imágenes de satélite y resultados de modelos de predicción meteorológicos comunes. El objetivo principal de este trabajo es adoptar un marco de trabajo que permita, de forma distribuida y en tiempo real, la integración de datos relacionados con eventos meteorológicos; incluyendo la capacidad de leer, modificar, filtrar e integrar los datos de entrada y hacer procesos de minería de datos en tiempo real. Los autores introducen la descripción de un proyecto llamado ADMIRE (Advanced *Data Mining* and Integration Research for Europe) el cual será tomado como base para su propósito.

Fortalezas

- Presentan un esquema completo que incluye el tratamiento necesario para los datos, desde el momento que se generan hasta el momento en el que se puede extraer conocimiento de ellos.
- Dan una importancia marcada a la necesidad de entender los fenómenos meteorológicos que están detrás de los datos recolectados y la influencia que estos tienen para diferentes actividades de la sociedad como los sistemas de transporte, la agricultura, la prevención de desastres, entre otros.
- Presentan el proyecto ADMIRE el cual provee un marco de trabajo completo para el análisis e integración de datos y que puede ser usado por la comunidad académica para sus investigaciones.

Limitaciones

- El trabajo presentado es un proyecto, en el cual no se tienen aún datos concretos respecto a la viabilidad de su implementación y a los resultados que pueda llegar a alcanzar.

Oportunidades

- Se va a revisar el proyecto ADMIRE para determinar si se puede adaptar al caso de estudio particular de esta tesis y evaluar los resultados que se obtengan.

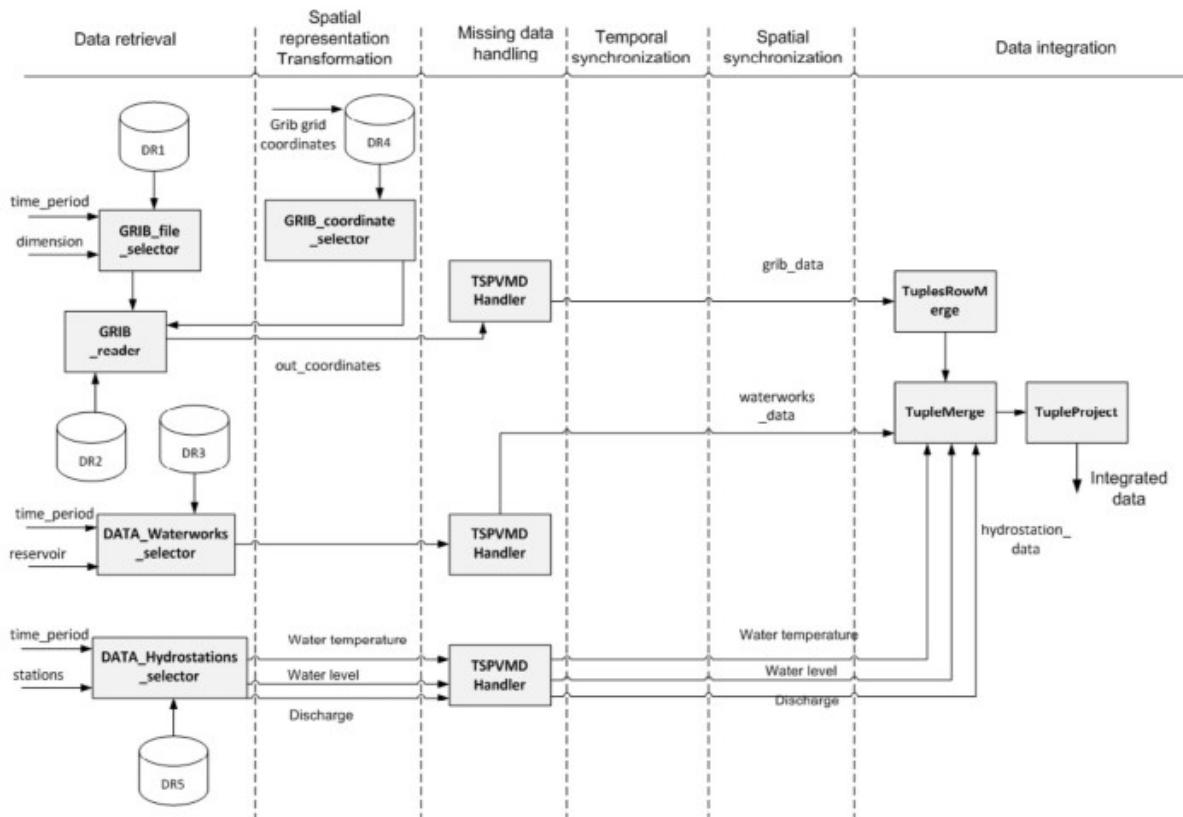


Figura 3-6: Integración de datos y procesos de minería usando datos ambientales distribuidos. Tomado de (Bartok et al., 2010)

- ADMIRE (Advanced *Data Mining* and *Integration Research for Europe*) es un proyecto desarrollado desde el año 2007 al 2011 por un grupo de colaboración con investigadores y desarrolladores de varias universidades europeas. Combina estrategias y tecnologías para la creación de una plataforma de descubrimiento de conocimiento que incluye el acceso a datos, el pre-procesamiento e integración, la minería de datos, el análisis estadístico, el pos-procesamiento, transformación y entrega. Entre las características de la plataforma ADMIRE están: DISPEL, un potente lenguaje Java para la descripción de flujos de trabajo con intensidad de datos; un motor de ejecución streaming para eliminar los cuellos de botella de datos; descripciones semánticas de elementos del flujo de trabajo basadas en una red común de ontologías y algunas herramientas de programación visual basadas en la plataforma eclipse. La plataforma ADMIRE incluye una arquitectura (ver Figura 3-7), un marco de trabajo y un lenguaje. Se ha aplicado con éxito a un

escenario con datos ambientales para la realización de predicciones hidrológicas. El escenario utiliza varios conjuntos de datos, que se distribuyen geográficamente (Simo, Habala, Tran, Krammer, & Hluchy, 2011). La integración y minería de datos se hace como un proceso iterativo, los usuarios pueden intentar usar algunos conjuntos de datos y métodos de minería, examinar los resultados y luego cambiar los conjuntos de datos y / o métodos de extracción para obtener mejores resultados.

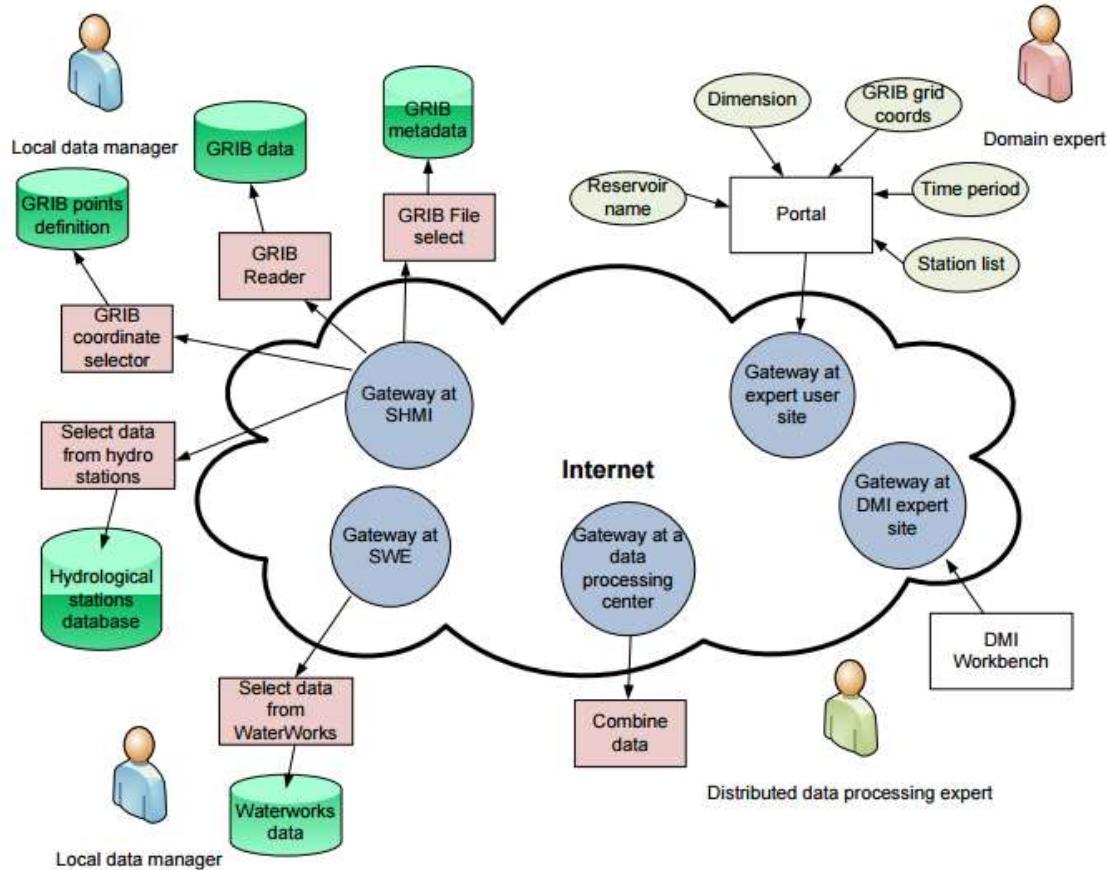


Figura 3-7: Arquitectura ADMIRE. Tomado de (Simo et al., 2011)

Fortalezas

- Ofrece un apoyo para la toma de decisiones, por medio de un mejor uso de los datos en procesos de negocio o investigación.
- Brinda una potente metodología y un lenguaje común para el desarrollo de soluciones de integración de la minería de datos y conjuntos de datos cada vez más grandes y complejos.

- Presenta una serie de herramientas y métodos para separar los procesos intensivos de datos de su implementación, apoyando con esto el análisis de datos.

Oportunidades

- El lenguaje desarrollado en el proyecto puede ser probado para el tratamiento de los datos del caso de estudio particular de esta tesis.
- (Song et al., 2015) presentan el diseño, implementación y evaluación de HaoLap (OLAP basado en *Hadoop*), un sistema OLAP para *Big Data*. Los autores proponen un modelo de codificación de las dimensiones y un algoritmo de recorrido para el tratamiento de las operaciones de jerarquía de las dimensiones. El rendimiento de este modelo fue comparado con Hive, *HadoopDB*, HBaseLattice y Olap4Cloud, obteniendo buenos resultados. Las contribuciones de este trabajo son principalmente, (1) HaoLap adopta muchos enfoques para optimizar el rendimiento de OLAP, por ejemplo un modelo multidimensional simplificado para mapear dimensiones y medidas, (2) los autores usaron algoritmos de partición y linealización para almacenar los datos y una estrategia de selección por fragmentos para el filtrado de datos, (3) desarrollaron un algoritmo OLAP y un algoritmo de carga de datos en *MapReduce*, HaoLap almacena las dimensiones en metadatos y las medidas en un HDFS sin introducir almacenamiento duplicado. El sistema incluye cinco componentes: cluster *Hadoop*, un servidor de metadatos, un Nodo de trabajo, interfaz de servicios OLAP y un cliente OLAP. En la Figura 3-8 se presenta la arquitectura del sistema, cabe anotar, que el componente de ETL mostrado en la figura debe ser definido por el usuario para extraer los datos desde fuentes específicas.

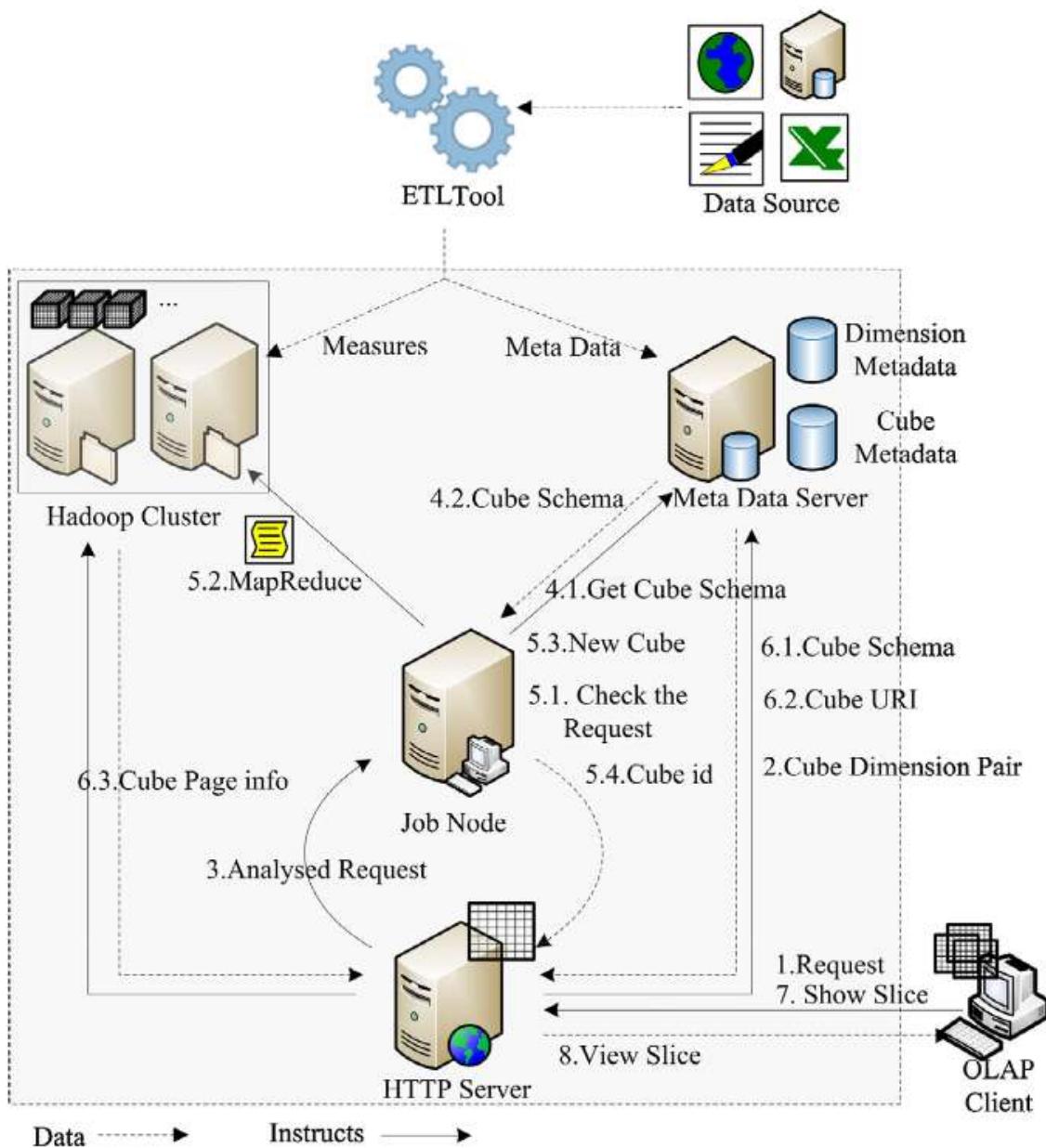


Figura 3-8: Arquitectura del sistema HaoLap. Tomado de (Song et al., 2015)

Fortalezas

- HaoLap toma las ventajas del modelo de datos especial, simplificando el modelo dimensional en el que se basa, manteniendo los procesos OLAP simples y eficientes.

- HaoLap adopta métodos de partición y selección por fragmentos para aumentar la velocidad de OLAP y dar cabida a entornos con *Big Data*.

Limitaciones

- El sistema no considera un módulo para la realización de ETL, por lo cual esto debe ser considerado previamente por el usuario y enganchado a HaoLap.

Oportunidades

- El código fuente de HaoLap está disponible en <https://github.com/MarcGuo/HaoLap> para hacer pruebas con fuentes de datos propias y comprobar sus ventajas y/o falencias.

3.4 Síntesis de los trabajos del segundo grupo

Para hacer la síntesis de este grupo de trabajos que presentan el diseño y/o implementación de modelos para el almacenamiento de grandes volúmenes de datos y procesos de análisis, se presenta en la Tabla 3-2 un comparativo que considera algunas características que se consideran primordiales y se detallan a continuación. Presencia de un esquema de almacenamiento definido, es decir, se expone el diseño de un modelo de datos claro y que permite estructurar las fuentes de datos y trabajar sobre él. Preprocesamiento y ETL sobre los datos, se garantiza que los datos con los que será alimentado el modelo han sido sometidos a una limpieza previa con el fin de eliminar errores y ruido. Extracción y análisis de datos, a los datos almacenados se les hace algún tipo de tratamiento que permita la extracción de información y posterior análisis. Aplicación de técnicas de minería de datos y/u otras de inteligencia artificial, dentro del trabajo se describe la aplicación de alguna técnica en particular para el tratamiento y análisis de los datos. Se consideran diferentes fuentes de datos, el modelo es pensado para recibir datos de diferentes fuentes como por ejemplo bases de datos de históricos, mediciones de sensores, información de internet, datos no estructurados, entre otras. El signo de interrogación denota la falta de claridad para ese aspecto de acuerdo a lo presentado por los autores en el artículo.

Tabla 3-2: Síntesis del segundo grupo

Trabajo relacionado	Presencia de un esquema de almacenamiento	Extracción y análisis de datos	Aplicación		Pre-procesamiento, ETL sobre los datos
			de técnicas de minería de datos	Diferentes fuentes de datos y/o IA	
Duque et al. (2014)	SI	SI	SI	SI	SI
Bartok, et al. (2010)	?	?	SI	SI	?
Simo et al. (2011)	?	SI	SI	SI	SI
Song et al. (2015)	SI	SI	?	NO	NO

3.5 Tercer grupo: aplicación de técnicas para la predicción de comportamiento de variables climáticas

En este tercer grupo se presentan algunos trabajos relacionados con la aplicación de técnicas de inteligencia artificial, para la realización de predicciones de variables ambientales, principalmente de precipitación.

- (Sawale & Gupta, 2013) presentan un algoritmo basado en una red neuronal artificial para predecir la condición atmosférica en una localización y tiempo determinados, a partir de una serie de datos meteorológicos, que incluye las variables temperatura, humedad y velocidad del viento. Los autores emplearon una red neuronal de tipo Back Propagation (BPN) para el modelado inicial. Los resultados obtenidos con esta red BPN alimentan a una red Hopfield. El modelo basado en la combinación de una red BPN y una red Hopfield fue probado en un conjunto de datos que comprende mediciones de 3 años de datos meteorológicos (un total de 15.000 registros) con atributos como temperatura, humedad y velocidad del viento. Los resultados obtenidos muestran que el error de predicción es bajo.

El objetivo principal de este trabajo se encuentra en la aplicación de minería de datos predictiva para la extracción de patrones interesantes (no triviales, implícitos, desconocidos y potencialmente útiles) a partir de datos meteorológicos.

Fortalezas

- Se construyó un modelo con un enfoque combinado, a partir de una RNA tipo back propagation y una red Hopfield, lo anterior para lograr una mejor predicción.
- Los datos fueron limpiados y filtrados previamente, lo cual garantiza que al llegar a la fase de entrenamiento ya se contaba con datos confiables.
- Así mismo, los datos con los que se entrenó la red fueron normalizados, lo anterior permite eliminar ruido y contrarresta los problemas que se pueden generar por la diferencia de magnitudes y unidades de las variables tratadas.

Limitaciones

- Los resultados obtenidos no son presentados claramente en el artículo, solo se describen cualitativamente, pero no se hace claridad sobre el porcentaje de acierto o error de las predicciones.
- Los datos analizados corresponden a un *dataset* de un periodo de tiempo no muy largo, lo cual hace que no se pueda garantizar que el modelo y los resultados obtenidos puedan ser replicados a una zona mayor.

Oportunidades

- Se puede probar el uso de este modelo que combina dos tipos de RNA para un *dataset* propio y de mayor volumen para evaluar los resultados y la pertinencia de las predicciones generadas.
- (Duque Mendez, Orozco Alzate, & Hincapié, 2011) presentan los resultados de un proceso de tratamiento de datos meteorológicos con técnicas de minería de datos para la extracción de conocimiento y para la generación de predicciones preliminares de la variable temperatura promedio en un periodo de tiempo de 24

horas. Se emplearon datos recolectados por algunas de las estaciones de monitoreo climático de la ciudad de Manizales para un periodo de dos años (2005 – 2007). Los datos fueron sometidos a un proceso de limpieza previo a su análisis. El análisis de los datos se hizo por medio de la herramienta WEKA, se aplicaron métodos de agrupamiento y métodos de clasificación; finalmente por medio de estos últimos se hizo la predicción de la variable temperatura. La Figura 3-9 presenta los algoritmos y técnicas usadas. Los resultados en la predicción de esta variable fueron satisfactorios, así mismo se lograron identificar relaciones entre otras variables.

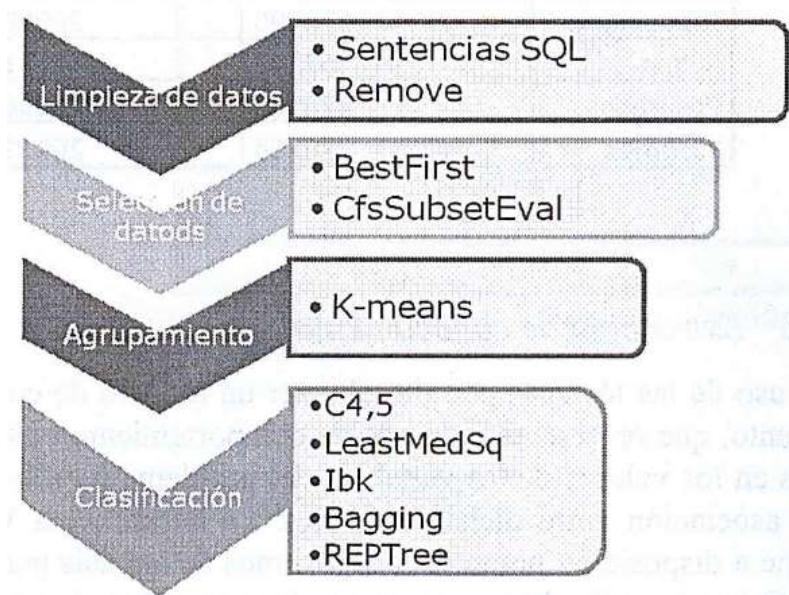


Figura 3-9: Algoritmos y técnicas utilizadas. Tomado de (Duque Mendez et al., 2011)

Fortalezas

- Se tuvo en cuenta la necesidad de hacer una limpieza inicial a los datos para eliminar valores atípicos.
- Los resultados presentan un acercamiento inicial a la predicción de variables climáticas.

Limitaciones

- Las series de tiempo trabajadas son cortas y presentan baches de tiempo sin datos, lo cual puede ser una limitante a la hora de construir un modelo para reconocimiento de patrones.

- No se aclara si los algoritmos que encontraron predicciones positivas para una de las estaciones fueron aplicados a otras estaciones y produjeron resultados similares.

Oportunidades

- Se puede seguir ampliando el modelo propuesto con la automatización de procesos como la limpieza de datos inicial y con la inclusión de otras estaciones de monitoreo.
- (Beltrán-Castro, Valencia-Aguirre, Orozco-Alzate, Castellanos-Domínguez, & Travieso-González, 2013) proponen la utilización del principio de descomponer y ensamblar para la generación de una metodología de predicción de la precipitación. La técnica de descomposición utilizada es la descomposición modal empírica de ensamble (EEMD) y como herramienta de predicción adoptan una red neuronal FNN (Feed Forward Network). La metodología EEMD tiene como objetivo hacer frente a la tarea de predecir una señal compleja, descomponiéndola inicialmente en partes más simples que se pueden modelar con mayor precisión, y luego agregando o uniendo esos resultados en un pronóstico final de la señal original. La Figura 3-10 muestra la metodología seguida por los autores. Para determinar las variables de entrada a cada red FNN se empleó la función de autocorrelación parcial (PACF, por sus siglas en inglés), en busca de los retardos en los que la PACF estuviese fuera del intervalo de confianza, y el número de nodos ocultos se ajustó a $2p + 1$, donde p es el número de entradas. El modelo fue probado con datos de una estación meteorológica de la ciudad de Manizales y se obtuvo que se presentan mejoras en la predicción obtenida usando este método frente a una predicción obtenida con una única red FNN.

Fortalezas

- Emplear el principio de descomponer y ensamblar por medio de la técnica EEMD, mejoró la predicción frente a la obtenida por una única red FNN.

- Se trabajaron tres criterios de evaluación de la predicción, el error absoluto medio (MAPE), el error cuadrático medio (MSE) y el error absoluto medio (MAE)

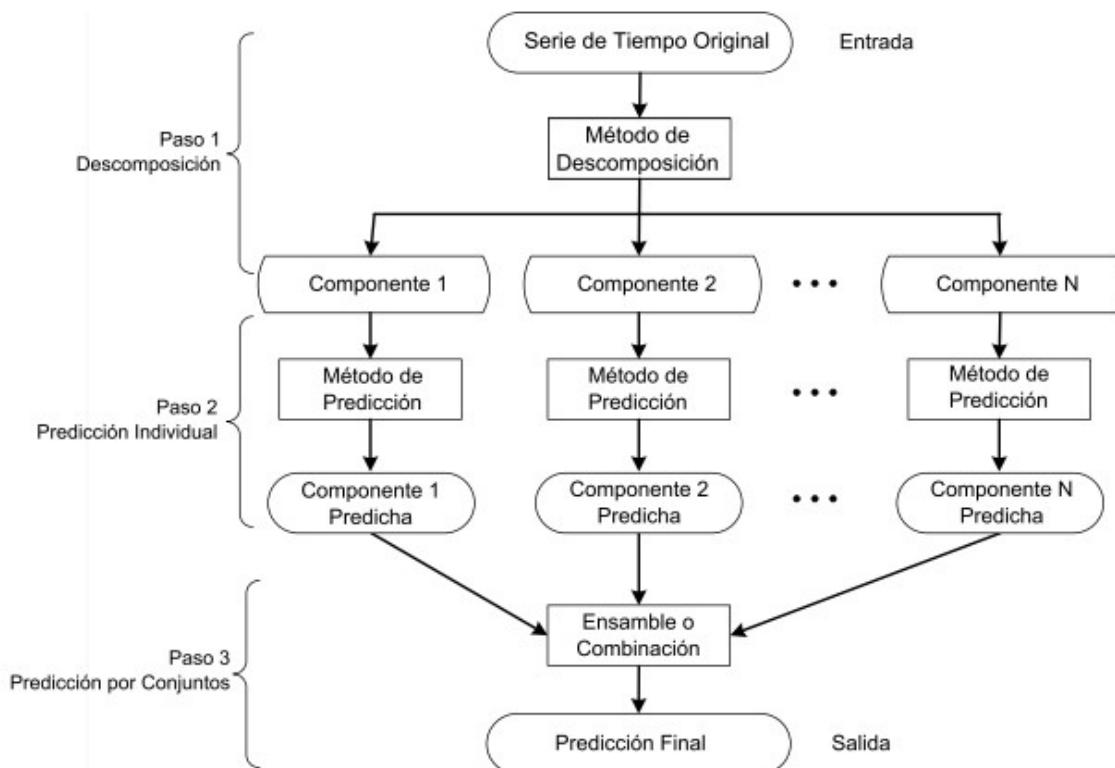


Figura 3-10: Esquema metodológico. Tomado de (Beltrán-Castro et al., 2013)

Limitaciones

- No se tienen en cuenta otras variables meteorológicas como atributos explicativos para hacer la predicción, solo se toman valores de la precipitación.

Oportunidades

- Se pueden realizar nuevas pruebas incluyendo otros *datasets* para hacer más fiables los resultados y efectuar comparaciones con técnicas de predicción tradicionales probando el esquema a largo plazo.

- (Abhishek, Kumar, Ranjan, & Kumar, 2012) proponen la utilización de una red neuronal artificial para la predicción de la precipitación promedio mensual en una zona de la India caracterizada por el sistema climático de tipo Monzón. La red es implementada en Matlab utilizando los paquetes Nntool (open network/data manager) y Nftool (Neural network fitting tool). Para predecir la precipitación se utilizan como variables explicativas la humedad promedio y la velocidad del viento promedio para ocho meses de 50 años, desde 1960 a 2010. Los datos son normalizados utilizando la media y desviación estándar. Los experimentos se hacen con tres tipos de redes diferentes (Feed Forward with Back –Propagation, Layer Recurrent, Cascaded feed Forward back Propagation) y se plantean ocho casos o configuraciones para cada tipo de red, estas configuraciones son obtenidas variando la función de entrenamiento, la función de aprendizaje y el número de neuronas. Luego se comparan los resultados obtenidos con cada una, encontrando que el tipo de red que mejores resultados obtuvo fue Feed Forward with Back –Propagation y que al aumentar el número de neuronas en la capa oculta decrece el MSE.

Fortalezas

- Utilizan variables explicativas (Humedad promedio y la velocidad del viento promedio).
- Mejoran la predicción al aumentar el número de neuronas. Para su caso de estudio la configuración en Matlab que mejores resultados da es: LEARNGDM como función de aprendizaje, TRAINLM como función de entrenamiento.

Limitaciones

- Para el caso de estudio presentado, se utilizan solo los 8 meses del año en los que se tiene certeza de que se presentarán los eventos de precipitación, el tipo de clima monzón se diferencia en gran medida del clima de regiones ecuatoriales.

- (Luk, Ball, & Sharma, 2000) presentan un trabajo que tuvo como objetivo identificar los datos espacio-temporales necesarios para lograr pronósticos de precipitación más precisos y a corto plazo. Se empleó una red neuronal artificial para reconocer los patrones de lluvia históricos a partir de una serie de mediciones realizadas en la cuenca de estudio y luego poderlos reproducir para nuevos eventos de lluvia. Para lograr este objetivo, los autores realizaron pruebas de comparación sobre la precisión de los pronósticos realizados por las redes neuronales configuradas con diferentes órdenes de retraso y número de entradas espaciales. Para el desarrollo de las RNAs con configuraciones alternativas, se llevaron a cabo varios entrenamientos con el fin de alcanzar un nivel óptimo y lograr una buena generalización de los datos. Finalmente el estudio encontró que se obtienen predicciones más precisas cuando las RNAs cuentan con un número óptimo de entradas, así mismo, la red que empleó un menor retraso obtuvo un mejor rendimiento.

Fortalezas

- En el estudio se planteó también el uso de información de medidores cercanos (vecinos) y los mejores resultados se dieron con 8 vecinos y considerando solo la precipitación en el t-1, teniendo en cuenta que manejan mediciones cada 15 minutos.

Limitaciones

- No se miran más variables aparte de la precipitación en mediciones anteriores.
- No se hace referencia a la configuración de la RNA (algoritmo de entrenamiento, de aprendizaje), ni al software utilizado para las pruebas
- (Sahai, Soman, & Satyan, 2000) introducen un modelo para el pronóstico de la precipitación mensual del verano monzón y de escalas de tiempo estacionales con el uso de red neuronal artificial, uno de los principales aportes de este trabajo es la mejora en el cálculo del bias usado por la red. Los autores determinan, por medio de estadísticas de verificación, que las redes neuronales artificiales son buenos predictores de las precipitaciones medias estacionales y mensuales en la India,

utilizando sólo series temporales de lluvia como entradas. Con la utilización de datos de solo cinco años anteriores de los valores de las precipitaciones medias mensuales y estacionales como insumo para predecir los valores del próximo año, se lograron buenos resultados para plazos de mucho más tiempo (8 meses) en comparación con las predicciones realizadas con métodos estadísticos convencionales. Los datos empleados para las pruebas son divididos en conjuntos de entrenamiento y validación, dando con esto mayor confiabilidad a los resultados.

Fortalezas

- Por la naturaleza del fenómeno estudiado (lluvia de verano monzón) y gracias a la optimización de la asignación del bias de la RNA se pueden hacer predicciones a largo plazo (8 meses a 1 año)

Limitaciones

- Solo se toma la precipitación como variable para alimentar el modelo de predicción.
- Liu, Li, & Dillon (2001) desarrollan una mejora a la técnica del clasificador Naïve Bayes y exploran el uso de Algoritmos Genéticos para la selección de atributos de entrada en problemas de clasificación y predicción de precipitación. Se plantean dos esquemas simples para construir un conjunto de datos adecuado para mejorar el rendimiento de la predicción. En el esquema I utilizan todos los parámetros de entrada básicos para la predicción de la precipitación (meses, temperatura, precipitación, presión atmosférica, velocidad del viento, direcciones de la rosa de los vientos, humedad, brillo solar, evaporación y punto de rocío). En el esquema II utilizan un subconjunto óptimo de variables de entrada que han sido seleccionadas por el Algoritmo Genético. La adopción de Algoritmos Genéticos para la selección de variables de entrada realizada en este estudio, demuestra que existe una viabilidad en la reducción de la complejidad del problema de predicción con un rendimiento comparable o ligeramente mejor. Los autores ratifican que la predicción del tiempo seguirá siendo un importante reto científico para futuras investigaciones.

Fortalezas

- Los autores consideran una técnica para hacer la selección de las variables de entrada que pueden ser buenas predictores de la salida.
- Se hace una clasificación previa de la precipitación (lluvia - no lluvia) y luego una categorización de la lluvia como ligera, moderada y fuerte; con lo anterior se puede facilitar el proceso de predicción ya que el problema es dividido en dos fases.

Limitaciones

- El número de variables explicativas utilizadas como parámetros de entrada puede no ser suficiente para capturar todas las características necesarias para el período de predicción de 24 horas, ya que cualquier cambio significativo de las condiciones climáticas pueden tener lugar durante o después de este período.

Oportunidades

- Se puede utilizar la idea de generación y reducción de variables para facilitar el descubrimiento de nuevas heurísticas de predicción.
- Tener fuentes de atributos de entrada adicionales, como por ejemplo datos de mediciones en otros puntos cercanos puede ser útil para mejorar el rendimiento en las predicciones.
- En el estudio presentado por (Valverde Ramírez, de Campos Velho, & Ferreira, 2005) se hace uso de una Red Neuronal de tipo feedforward para la predicción cuantitativa de la precipitación diaria en sitios específicos de la región de São Paulo, Brasil. La red utiliza el algoritmo de aprendizaje de propagación resiliente y es alimentada con datos de variables meteorológicas como temperatura potencial, componente vertical del viento, humedad específica, temperatura del aire, entre otras, todas del período 1997-2002. Los resultados arrojados por la red para la predicción son comparados con una regresión lineal múltiple por medio de análisis estadístico, con el fin de evaluar la habilidad de pronóstico de lluvia sobre la región estudiada. Los resultados muestran que los pronósticos de la RNA fueron superiores a los obtenidos por el modelo de regresión lineal. Con el fin de mejorar

la capacidad de predicción del modelo, se necesitan datos adicionales que lo alimenten, como por ejemplo imágenes de satélite y de series de tiempo con observaciones de mucho más tiempo

Fortalezas

- El modelo con una RNA es una herramienta de mapeo no lineal, logrando con esto que sea potencialmente más adecuado para la predicción de la lluvia, ya que esta variable tiene una naturaleza no lineal.

Limitaciones

- Para la selección de las variables asociadas al modelo, se usa un estudio previo y conocimiento que se tiene a priori del caso de estudio, no se describe el uso de ninguna técnica para la selección de atributos.
- (Toth, Brath, & Montanari, 2000) realizan una comparación de técnicas de análisis de series temporales para la predicción de precipitación a corto plazo. Las entradas a los modelos son datos de precipitación en períodos anteriores. Las predicciones obtenidas con los métodos comentados son agrupadas en un modelo conceptual de lluvia escorrentía y se implementa en un caso de estudio de los montes Apeninos, Italia. Las técnicas empleadas son modelos autorregresivos de media móvil (ARMA), redes neuronales artificiales (RNA) y el método de vecinos cercanos no paramétrico (K-NN).

Fortalezas

- En general, el estudio indica que las técnicas de análisis de series temporales consideradas proporcionan una mejora en la precisión de los pronósticos de inundación con respecto al uso de los enfoques de predicción de precipitaciones intuitivos, heurísticos.
- El estudio concluye que las observaciones de precipitaciones anteriores por si solas no son suficientes para predecir eventos futuros, incluso a muy corto plazo.

Limitaciones

- Los resultados muestran que las predicciones son satisfactorias a muy corto plazo (1 a 6 horas) lo cual implica que no se tienen predicciones con un mayor rango de tiempo que permitan a las entidades de control o monitoreo ambiental hacer alertas a la comunidad para casos donde se requiera, por ejemplo una evacuación.
- (Dhanya & Nagesh Kumar, 2009) hacen uso de un algoritmo de minería de datos, Patrón Frecuente, para revisar reglas de asociación difusas (negativas y positivas) entre los índices ENSO y EQUINOO respecto al fenómeno de la precipitación de verano Monzón o verano húmedo (ISMR) en dos regiones homogéneas de la India. Las reglas se definen usando un conjunto de datos de entrenamiento del período 1958-1999 y validados con datos del período 2000-2006. A partir de las reglas construidas se consiguen unas salidas difusas que luego se convierten en salidas definidas utilizando un algoritmo de conteo ponderado. Se obtiene que la variabilidad de la precipitación del verano monzón de los últimos años es bien modelada por esta técnica, demostrando ser eficaz incluso cuando la relación estadística lineal entre los índices es débil.

Fortalezas

- La generación de las reglas de asociación es una buena propuesta para considerar en el caso de estudio de predicción meteorológica, incluyendo el conocimiento de expertos en eventos de este tipo.
- El método de las reglas de asociación difusas probado mejora los resultados frente a otros métodos de relación estadística lineal.

Limitaciones

- No se presenta una comparación del rendimiento del método empleado frente a otros aplicados en estudios previos.
- (Kusiak, Wei, Verma, & Roz, 2013) hacen la construcción de tres modelos para el pronóstico de la precipitación a partir de la evaluación de cinco algoritmos de minería de datos: redes neuronales, arboles aleatorios, árboles de regresión y

clasificación, máquinas de vectores de soporte y los k-vecinos más cercanos. Los modelos se alimentan con datos recolectados por medio de un radar. Los resultados son comparados contra un modelo lineal para evaluar la precisión en las predicciones.

Los datos fueron procesados previamente para garantizar que estuviesen en la misma frecuencia y para eliminar los valores atípicos y los valores negativos. Las predicciones realizadas son a corto plazo, el horizonte de tiempo más largo de predicción aceptable fue de 120 minutos.

Fortalezas

- Se hace una evaluación de cinco algoritmos de minería de datos, encontrando que el que presenta mejores resultados en las predicciones es la red neuronal (multilayer perceptron)

Limitaciones

- El horizonte de tiempo más largo en el que se obtiene una buena predicción es solo de 120 minutos.

Oportunidades

- Para futuras investigaciones, se recogerán datos de otras regiones con el fin de validar y mejorar el enfoque propuesto
- (Maino, Uzal, & Granitto, 2014) En este trabajo se aborda el problema de predicción de series temporales caóticas obtenidas de sistemas dinámicos no lineales determinísticos. Se presenta una técnica basada en redes neuronales profundas y se evalúa su rendimiento frente a las redes neuronales convencionales. Se considera en particular el problema de predicción para múltiples horizontes utilizando dos estrategias: el uso de redes de salida-múltiple frente a redes convencionales de salida-simple. Los resultados sobre las series temporales consideradas muestran un mejor desempeño de las arquitecturas profundas de salida simple.

Fortalezas

- El reemplazo de la función de activación sigmoidea utilizada convencionalmente, por unidades de activación ReLUs (Rectified Linear Units).
- Con este tipo de unidad de activación, la cantidad de épocas de backpropagation, es aproximadamente 6 veces menor que al usar unidades sigmoideas.
- Al considerar redes profundas con igual cantidad de parámetros que las redes no-profundas, no representó una diferencia considerable en cuanto al tiempo de ejecución, por lo tanto vale la pena considerar el uso de estas arquitecturas para la tarea de predicción.

Limitaciones

- Las series temporales trabajadas en esta investigación difieren de las series meteorológicas, por lo cual no se puede asegurar que se presenten resultados satisfactorios al replicar estas configuraciones.
- En este trabajo (James N.K Liu, Hu, You, & Chan, 2014) se concentraron en el desarrollo de una nueva aplicación del enfoque de Redes Neuronales Profundas (DNN por sus siglas en inglés de Deep Neural Network) o también más conocido como Aprendizaje Profundo (DL por sus siglas en inglés de Deep Learning). Este enfoque recientemente ha sido un tema muy popular en la comunidad investigativa. Con el uso de DNN, un conjunto de datos originales se puede representar en un nuevo espacio de características por medio de la aplicación de algoritmos de aprendizaje automático (learning machine) y de modelos inteligentes, con la posibilidad de obtener un mejor rendimiento en el espacio de características "aprendido". Los científicos han logrado resultados alentadores mediante el empleo de DNN en algunos campos de investigación que incluyen la visión por computador, el reconocimiento de voz, la programación lingüística y el procesamiento de información biológica. Los autores plantean que al contar con buenas características de aprendizaje, el DNN puede llegar a ser un enfoque más universal y tener potencial en otros dominios de datos y para otro tipo de problemas; por ello presentan una investigación inicial sobre la aplicación de DNN para hacer frente a problemas en series temporales meteorológicas. En la investigación aplican DNN para procesar datos masivos que involucran millones de registros ambientales

proporcionados por el Observatorio de Hong Kong (HKO). Las características obtenidas se utilizan para predecir el cambio de clima en las próximas 24 horas, tomando cuatro variables: temperatura, punto de rocío, presión media a nivel del mar (MSLP) y velocidad del viento.

Fortalezas

- Los resultados muestran que el DNN es capaz de ofrecer un buen espacio de características para los conjuntos de datos de clima y es también una herramienta potencial para la fusión de características en los problemas de series de tiempo.
- Se muestra que existe una relación estrecha entre el número de capas ocultas y la exactitud de la salida.
- DNN se puede combinar con muchos otros modelos, y las características aprendidas se pueden emplear para mejorar el rendimiento de dichos modelos.

Limitaciones

- La relación entre el número de capas ocultas y la exactitud de la salida debe ser cuantificada; también, el número de neuronas ocultas se debe optimizar.

Oportunidades

- Tratar de emplear este modelo con datos meteorológicos más complejos como por ejemplo datos de precipitación.
- (Grover, Kapoor, & Horvitz, 2015) presentan un modelo de predicción del tiempo el cual hace sus predicciones por medio de consideraciones de la articulación e influencia de variables climáticas esenciales, para el caso de estudio tomaron datos de vientos, temperatura, presión y punto de rocío de los años 2009 a 2015 observados en 60 puntos de los Estados Unidos. El modelo presentado cuenta con un nuevo enfoque híbrido que tiene componentes discriminativos y generativos para realizar inferencias espacio-temporales sobre el clima. La arquitectura propuesta combina un predictor “bottom-up” para cada variable individual con una red de creencia profunda “top-down” que modela las relaciones estadísticas

conjuntas. El otro componente clave en el marco de trabajo es un kernel basado en datos, que está dado sobre la base de una función de similitud que se aprende de forma automática a partir de los datos. El kernel es usado para asignar dependencias de largo rango a través del espacio y para asegurar que las inferencias respetan las leyes naturales. El modelo es llamado “Deep Hybrid Model” y los resultados obtenidos por este son comparados con otro modelo propuesto en un trabajo previo por otro autor y con el modelo usado por la Administración Nacional Oceánica y Atmosférica (NOAA) de los Estados Unidos, mostrando un mejor rendimiento que estos.

Fortalezas

- Se proporciona un procedimiento de inferencia eficiente y que permite optimizar el modelo predictivo de acuerdo con fenómenos de gran escala.
- Se evalúan los métodos con una serie de experimentos que dejar ver el rendimiento y el valor de la metodología.
- Los experimentos muestran que la nueva metodología puede proporcionar mejores resultados frente a puntos de referencia como el NOAA e investigaciones recientes.

Limitaciones

- Las predicciones realizadas suelen ser a muy corto plazo
- No se tienen en cuenta eventos meteorológicos más complejos como por ejemplo la precipitación.

Oportunidades

- A través de la integración de datos adicionales en el modelo híbrido se podría revisar la posibilidad de lograr predicciones a más largo plazo.

3.6 Síntesis del tercer grupo

Para hacer la síntesis de este último grupo de trabajos, los cuales presentan modelos para la predicción de variables meteorológicas, se muestra en la Tabla 3-3 un comparativo que considera las siguientes características: realización de un proceso de limpieza inicial a los

Capítulo 3

datos que alimentaron las pruebas de cada modelo; uso de técnicas de inteligencia artificial o de minería de datos para realizar la predicción; los conjuntos de datos que alimentan el modelo corresponden a datos meteorológicos y uso de otro tipo de enfoques recientes como el aprendizaje profundo (Deep Learning). El signo de interrogación denota la falta de claridad para ese aspecto de acuerdo a lo presentado por los autores en el artículo.

Tabla 3-3: Síntesis del tercer grupo

Trabajo relacionado	Limpieza inicial de datos	Uso técnicas de IA o MD	Tratamiento de datos meteorológicos	Uso de otros enfoques como Deep Learning	Lugar de aplicación	Escala temporal de la predicción
Sawale & Gupta (2013)	SI	SI	SI	NO		
Duque et al. (2011)	SI	SI	SI	NO	Manizales, Colombia	Diaria
Beltrán-Castro et al. (2013)	?	SI	SI	NO	Manizales, Colombia	Diaria
Abhishek et al. (2012)	?	SI	SI	NO	Udupi (estado de Karnataka), India	Mensual
Luk et al. (2000)	NO	SI	SI	NO	Cuenca del río Parramatta (Sydney), Australia	Cada 15 minutos
Sahai, et al. (2000)	?	SI	SI	NO	India	Mensual
Liu et al. (2001)	SI	SI	SI	NO	Hong Kong	Diaria

Trabajo relacionado	Limpieza inicial de datos	Uso técnicas de IA o MD	Tratamiento de datos meteorológicos	Uso de otros enfoques como Deep Learning	Lugar de aplicación	Escala temporal de la predicción
Valverde Ramírez et al. (2005)	NO	SI	SI	NO	São Paulo, Brazil	Diaria
Toth et al. (2000)	NO	SI	SI	NO	Montes Apeninos, Italia	1 a 6 horas
Dhanya & Nagesh Kumar (2009)	?	SI	SI	?	India	Mensual
Kusiak et al. (2013)	SI	SI	SI	NO	Oxford	120 minutos
Maino et al. (2014)	NA	SI	NO	SI	NA	NA
Liu et al. (2014)	SI	NO	SI	SI	Hong Kong	24 horas
Grover et al. (2015)	?	NO	SI	SI	EE.UU	6-12-24 horas

4. Modelo propuesto

En el capítulo que se despliega a continuación se presenta y describe el modelo genérico propuesto para la administración de *Big Data* y se particulariza en un modelo específico que permite su aplicación en el dominio de datos ambientales. En primera instancia se hará una identificación de cada una de las capas que hacen parte de la propuesta. Posteriormente se expondrá las tecnologías que podrían ser utilizadas en cada una de las capas del modelo y finalmente se introduce el modelo específico para los datos ambientales, describiendo las particularidades de cada capa.

4.1 Identificación de las capas asociadas al modelo de administración de *Big Data*

Para el modelado de soluciones *Big Data* es necesario tener en cuenta algunos aspectos o pasos a seguir. En primer lugar se debe 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. En segundo lugar, se necesita hacer un proceso de extracción y limpieza de los datos, es necesario tomar los datos de la fuente de origen, perifilar y limpiar los datos que se requieran, adecuándolos a las necesidades y aplicando filtrado o transformación que permitan garantizar la calidad de los mismos. En tercer lugar, se requiere estructurar los datos para su posterior análisis, esto incluye la creación de una estructura lógica para los sets de datos tratados, almacenar estos datos bajo el medio elegido y empezar algunos análisis para hallar relaciones, alternativas, patrones, etc. Por último, se debe iniciar el modelado de los datos, esto comprende el procesamiento de los datos por medio de la aplicación de algoritmos, de procesos estadísticos, técnicas de minería, entre otras. Como paso adicional y fundamental está la interpretación de los

resultados obtenidos, es necesario considerar si la solución se adecua a los datos y a los intereses en su tratamiento, si se aporta un resultado final que permita generar valor a partir de todo el proceso que se ha llevado a cabo.

Teniendo en cuenta lo anterior, para el desarrollo de una solución *Big Data*, se debe contar con un modelo que permita llevar a cabo los pasos descritos y que cuente con los componentes para cada una de las fases. En la Figura 4-1 se presenta el esquema con las capas que se propone.

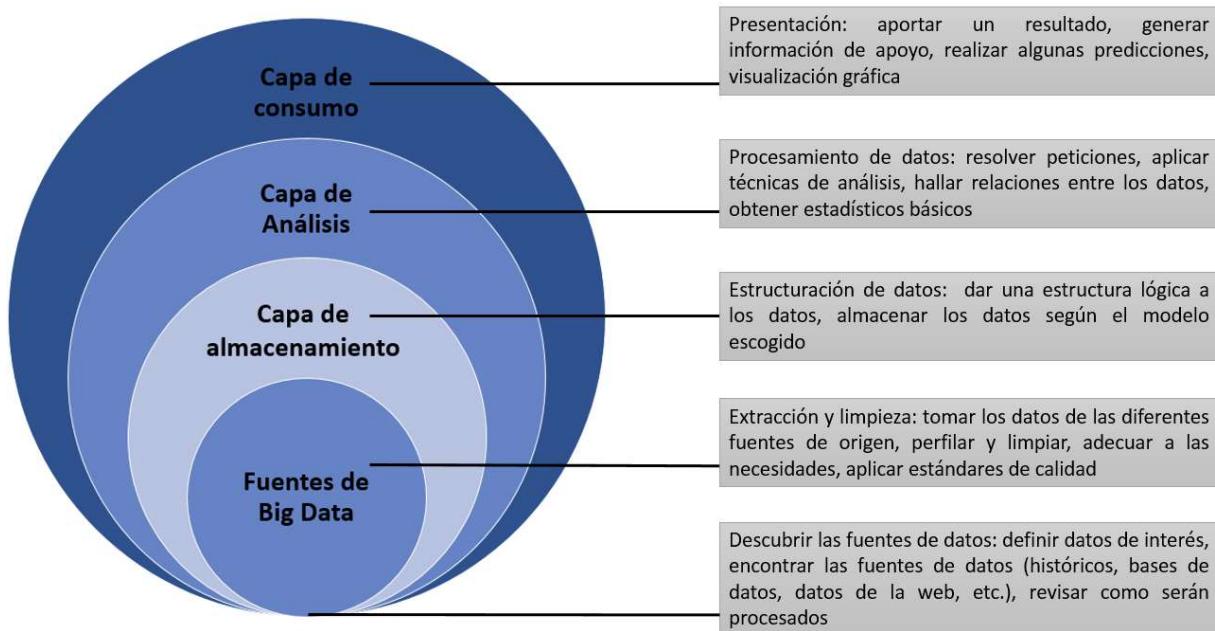


Figura 4-1: Capas del modelo genérico para la administración de *Big Data*. Elaboración propia.

A continuación, de manera detallada, se describirá cada una de las capas presentadas y la forma como pueden interactuar entre ellas.

4.1.1 Fuentes *Big Data*

Como se mencionó en el capítulo 2, existen diversas fuentes de *Big Data*, (George et al., 2014) las clasifica en cinco grupos, *Big Data* proveniente de datos públicos, datos privados, datos generados en la internet, datos de comunidades y datos autocuantificados. En

general, con el avance en la ciencia y la tecnología, casi todas las actividades de la vida diaria pueden ser monitoreadas, medidas y registradas; lo anterior genera que el número de datos que se producen, en un día y en un país específico, supere los miles de millones, todo esto constituye las fuentes de *Big Data*. En (Stephens et al., 2015) se muestran cuatro fuentes de *Big Data* que tienen un aumento considerable: datos astronómicos, datos genómicos, datos de Twitter y datos de YouTube; según la revisión de los autores, en cada uno de estos dominios, se proyectan al año la adquisición de 25 ZB de datos de astronomía, se generan de 0.5 a 15 billones de tweets, de 500 a 900 millones de horas de video se suben a YouTube y se generan 1 zetta-bases de datos genómicos.

Los datos pueden estar representados en diferentes fuentes, como por ejemplo, archivos planos, repositorios de datos estructurados (bases de datos), datos tomados de la web, flujos de datos, entre otros. Se pueden presentar problemas en su adquisición y pueden contener errores de recolección o reporte; por ejemplo, datos anormales que han sido tomados por sensores con problemas de funcionamiento. Por lo anterior es necesario contar con un proceso previo al almacenamiento, este proceso se reconoce como ETL y consiste en la toma y estandarización de la estructura de los datos, limpieza de datos extraños, llenado de datos faltantes y carga final. Existen herramientas para la realización de los procesos de ETL de forma genérica, algunos ejemplos se pueden encontrar en (Talend, 2016), (Pentaho, 2016) y (Javlin Inc., 2015). Sin embargo, al entrar a dominios de datos específicos se encuentra que las soluciones genéricas no siempre se ajustan a las necesidades particulares, por lo cual se requiere del conocimiento de expertos del tema para poder construir una solución que atienda las particularidades de los datos a tratar. En el capítulo 5 se describirá el esquema de ETL particular usado para el pre-procesamiento de los datos hidrometeorológicos antes de ser almacenados.

4.1.2 Capa de almacenamiento

La capa de almacenamiento puede ser vista como la base del modelo propuesto, en esta se toman los datos que han sido identificados en la fuente o fuentes de datos y que serán el objeto de la solución. En esta capa es necesario contar con una estrategia para la estructuración de los datos y elegir la plataforma que se utilizará para el almacenamiento de los mismos. Como se mencionó anteriormente, previo al almacenamiento en esta capa

es importante considerar el proceso de limpieza, extracción y carga – ETL – que se requiere para garantizar la calidad de los datos a almacenar.

Se pueden emplear en esta capa diferentes esquemas de almacenamiento, esquemas para datos estructurados y esquemas que soporten datos semi-estructurados o no estructurados. En la Tabla 4-1 se presenta un comparativo de las características de un esquema de almacenamiento en bases de datos relacionales frente a soluciones *Big Data*.

Tabla 4-1: Comparativo esquemas de almacenamiento. Adaptado de (Microsoft, 2014).

Característica	Sistema de bases de datos relational	Solución <i>Big Data</i>
Tipos de datos y formatos	Estructurados	Semi-estructurado y no estructurado
Integridad de datos	Alta – actualizaciones transaccionales	Depende de la tecnología usada - a menudo sigue un modelo de consistencia eventual
Esquema	Estático – requerido en escritura	Dinámico – opcional en escritura/lectura
Patrones de lectura y escritura	Lectura/escritura totalmente repetible	Lectura una vez, escritura repetible
Volumen de almacenamiento	Gigabytes a terabytes	Terabytes, petabytes y superior
Escalabilidad	Escalabilidad al tener hardware más potente	Escalabilidad con servidores adicionales
Procesamiento de datos distribuido	Limitado o nulo	Distribuido a través de clúster

Los sistemas de bases de datos relacionales usan típicamente modelos relacionales con esquemas predeterminados, tablas y relaciones definidas sobre datos estructurados. Existen también soluciones más flexibles, que soportan datos semi-estructurados y no estructurados. En la construcción de una solución *Big Data* se pueden emplear tanto esquemas tradicionales como los emergentes, ya que se busca atender a las necesidades y condiciones que presentan las diferentes fuentes de datos; de esta forma, se crean mecanismos complementarios que aumentan las capacidades para manejar los datos.

4.1.3 Capa de análisis

La capa de análisis es el centro del modelo, en esta capa se realiza el tratamiento de los datos. Se puede partir de la obtención de relaciones simples por medio de la aplicación de estadísticos básicos y pasar a otros tipos de análisis que incluya la aplicación de técnicas derivadas de la IA, minería de datos, machine learning o procesamiento distribuido. Siendo la capa que se comunica con las otras dos (almacenamiento y consumo) es de vital importancia que en esta se cuente con mecanismos que garanticen la capacidad de responder a las peticiones de los usuarios mediante la consulta a la capa de almacenamiento.

Al contar una variedad tan amplia de posibles técnicas a aplicar en la capa de análisis, la elección de estas, se fundamenta en el objetivo, en el tipo de datos y en los resultados que se vayan obteniendo de los análisis iniciales. La tarea de análisis de datos se soporta en técnicas específicas y sujetas al tipo de datos que se tienen o preparan para tal fin.

Puede que muchas de las técnicas no arrojen los resultados que se esperarían. En el caso particular de las técnicas de predicción y específicamente en algunos dominios con series de tiempo con comportamiento variable; el mecanismo para detectar cuál técnica se acopla mejor para la predicción es, además de la revisión de trabajos previos, el método de prueba y error; ya que, para muchos casos, no existen antecedentes sólidos que permitan determinar el éxito de alguna técnica o algoritmo específico previamente a su aplicación.

4.1.4 Capa de consumo

Esta capa tiene una importancia que radica en el contacto con los usuarios finales. Pues son ellos quienes podrán dar uso y aprovechar los resultados generados a partir de la capa de análisis, visualizar los datos de la capa de almacenamiento, detectar las necesidades de aplicación de nuevas técnicas o la necesidad de adquirir e integrar nuevas fuentes de datos.

La capa de consumo puede estar distribuida en diferentes medios, como aplicaciones o plataformas de presentación de informes (reporting), páginas web para la presentación de datos en tiempo real, sistemas de monitorización de redes, entre otras. Esta capa también debe garantizar que los datos puedan llegar a un mayor número de interesados, por esto, es bueno considerar principalmente la implementación de soluciones que permitan el acceso desde cualquier lugar y tiempo; lo anterior apunta al desarrollo de interfaces web.

La capa de consumo se conecta con la de análisis, pero en ocasiones puede tener también conexión con la de almacenamiento, porque se pueden requerir reportes que no requieran de un análisis específico, sino que simplemente se requiera de consultas de datos específicos, ya sea de un determinado periodo de tiempo o de una variable particular.

Si bien es cierto que la capa de consumo debe sintetizar en la mayor medida posible la información que tendrán que leer o visualizar los usuarios, expertos y demás interesados, es una ventaja contar con un volumen considerable de datos, ya que son el insumo principal de los análisis.

4.2 Tecnologías de apoyo para cada una de las capas del modelo

Tanto en las tecnologías asociadas a *Big Data*, como en las tecnologías para tratamiento de datos utilizadas en inteligencia de negocios o en soluciones de gestión de datos particulares, se encuentra un gran número de opciones a utilizar dentro de cada una de las capas del modelo propuesto. Nuevamente, la decisión en la elección de la tecnología depende directamente de la naturaleza del dominio y del tipo de datos que se tratarán. A continuación, se introducen algunas de las tecnologías que se han revisado y que pueden ser integradas a las capas propuestas en el modelo de gestión de *Big Data*.

Para la adquisición de datos desde las fuentes y para la realización de los procesos de ETL se encuentran herramientas comerciales y de código abierto que pueden ser bastante útiles. Estos procesos requieren de soluciones para la categorización, normalización remoción de duplicidad y redundancia en los datos. Cuando se maneja *Big Data* semiestructurado o no estructurado se suelen necesitar herramientas que hagan procesamiento por lotes (batch) para lograr producir salidas que puedan ser directamente visualizadas o combinadas con otros conjuntos de datos. La elección de las herramientas o tecnologías está asociada a los datos que se quieran explorar, analizar o visualizar.

Algunas herramientas de ETL que se pueden conseguir en el medio (algunas con versión free y otras solo bajo pago) son: Ab Initio, Beneti, BITool – ETL Software, CloverETL, Cognos Decisionstream (IBM), Data Integrator (herramienta de Sap Business Objects),

IBM Websphere DataStage (antes Ascential DataStage), Microsoft Integration Services, Oracle Warehouse Builder, WebFocus-iWay DataMigrator Server, Pervasive, Informática PowerCenter, Oxio Data Intelligence ETL full web, SmartDB Workbench, Sunopsis (Oracle), SAS Dataflux, Sybase, Syncsort: DMExpress, Opentext (antes Genio, Hummingbird), Pentaho Data Integration, Scriptella Open Source ETL Tool y Talend Open Studio.

Otra opción es el desarrollo de soluciones propias que atiendan directamente a los requerimientos de los datos a procesar y que se construyan en conjunto con los expertos en el dominio.

Las tecnologías a usar en la capa de almacenamiento dependen del tipo de datos que se quieran tratar, estructurados, semi-estructurados o no estructurados. Sin embargo, se puede considerar la posibilidad de combinar varias tecnologías o soluciones de almacenamiento dependiendo de si el problema considera fuentes de datos de diversa naturaleza. En el capítulo 2, se hizo la presentación de tecnologías para el almacenamiento de datos. Ahora, en la Figura 4-2 se hace una síntesis de algunas de estas, considerando su clasificación de acuerdo al tipo de datos y presentando algunos ejemplos de soluciones comerciales y libres.



Figura 4-2: Tecnologías para almacenamiento. Elaboración propia.

La capa de análisis, es la que presenta mayor variedad de tecnologías a partir de las cuales construir un esquema o generar observaciones, revisar relaciones y correlaciones entre los datos y descubrir tendencias o patrones. De los resultados obtenidos en esta capa depende que en la capa de consumo se pueda presentar información realmente útil. Como tecnologías para esta capa, además de las soluciones de análisis de datos que han surgido con el auge de *Big Data*, también se encuentran técnicas más sencillas, partiendo de consultas relacionales y pasando por el análisis multidimensional (ROLAP, MOLAP, HOLAP), la minería de datos (análisis predictivo, análisis descriptivo) y la aplicación de machine learning (árboles de decisión, reglas de asociación, algoritmos genéticos, redes neuronales artificiales, máquinas de vectores de soporte, algoritmos de agrupamiento, redes bayesianas). Para estas técnicas mencionadas existen proveedores de soluciones comerciales con un alto desarrollo y también está la posibilidad de hacer implementaciones propias pensando en *Big Data* y, partiendo de tecnologías como *Hadoop*, *Mahout* y *MapReduce*.

La capa de consumo está dedicada a los usuarios finales y reales conocedores del dominio de datos trabajado. En esta capa se debe brindar la posibilidad de aprovechar el conocimiento obtenido de la aplicación de análisis y también visualizar la información de predicciones o consultas y monitorear los datos, incluso en tiempo real. Para el soporte de esta capa, principalmente se construyen aplicaciones de “Front-end”, en las cuales para el usuario es totalmente transparente el procesamiento de los datos antes de aportar valor.

4.3 Modelo específico para la administración de *Big Data* ambiental

En el apartado 4.1 se hizo una descripción del modelo genérico propuesto para la administración de *Big Data* y en el apartado 4.2 se mencionaron las tecnologías de apoyo que se pueden incluir en cada una de las capas. Ahora, se presenta un modelo específico para la administración de *Big Data* ambiental, en el cual se considera de manera concreta, las fuentes de datos propias de este dominio. Así mismo, se estructura un esquema particular para el modelo específico, donde se reflejan las capas del modelo genérico. Este esquema se puede apreciar en la Figura 4-3. A continuación, se detallará cada uno de los componentes del esquema.

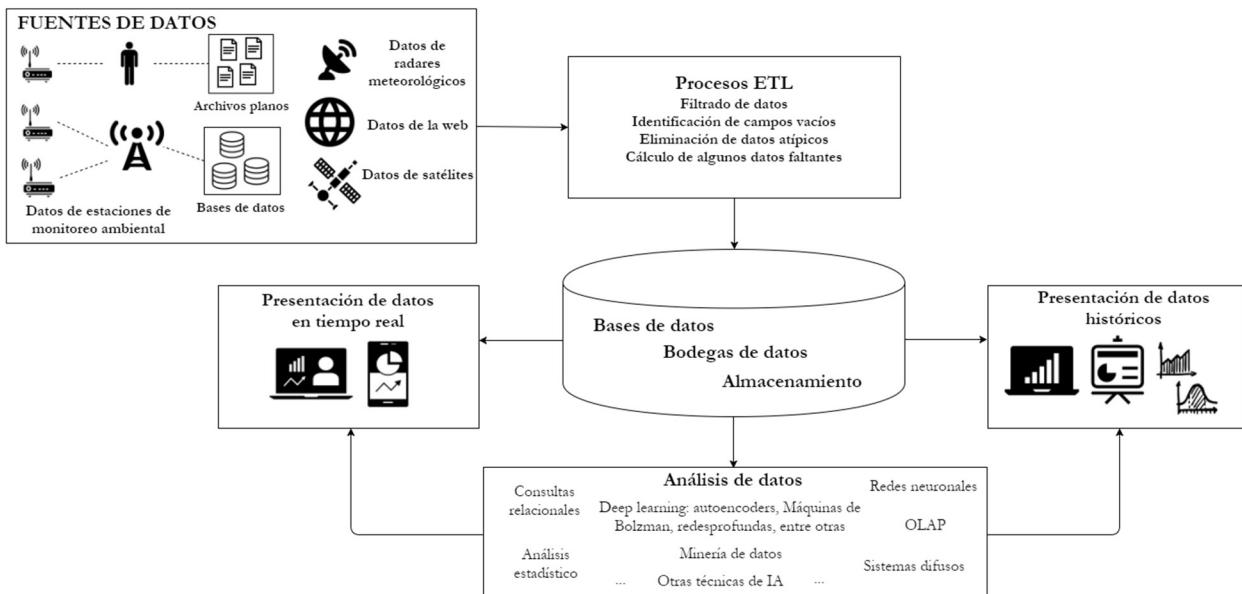


Figura 4-3: Modelo específico para la gestión de *Big Data* ambiental. Elaboración propia.

4.3.1 Fuentes de datos ambientales

Existen gran cantidad de fuentes de datos ambientales puesto que el número de variables ambientales es bastante amplio y se pueden extraer mediciones de estas, ya sea por medio de sensores o por la observación directa. Para el caso particular de esta tesis se considera un tipo de datos ambientales específicos, los correspondientes a mediciones de variables climáticas. Estas variables, según su naturaleza, son medidas por medio de sensores específicos o calculadas a partir de otras variables. Las redes hidroclimatológicas son una fuente constante de datos y a partir de las cuales se pueden generar numerosos trabajos investigativos y estudios puntuales (Vélez Upegui, Duque Méndez, Mejía Fernández, & Orozco Alzate, 2012).

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.

4.3.2 Proceso de ETL

Para realizar el proceso de ETL (extracción, transformación y carga), se requiere contar con la posibilidad de tomar diferentes fuentes de datos y también tener la capacidad de hacer un filtrado, tanto de detección como de corrección, garantizando con esto la integridad y consistencia de los datos que se almacenarán. Los datos pueden estar presentes en diferentes fuentes, como por ejemplo, archivos planos y repositorios de datos estructurados (bases de datos); estas fuentes pueden presentar problemas o errores a la hora de adquirir los datos. Por lo anterior es necesario contar con un proceso previo que puede ser llamado "traducción"; este proceso consiste en la toma y estandarización de la estructura de los datos, para luego proceder a la entrega de los mismos a la siguiente fase del proceso ETL.

Una fase posterior en el proceso, debe encargarse del filtrado y migración de los datos; considerando esto como la fase principal de todo el proceso. Acá se reciben los datos organizados en una estructura estándar, construida en la fase de traducción, y a partir de esto se realiza la tarea de filtrado; posteriormente, si es el caso, se deben hacer las correcciones predefinidas (aplicar filtros de corrección), las cuales han sido diseñadas gracias al conocimiento de expertos en el tema o dominio de datos. La segunda parte de la fase corresponde a la migración final de los datos hacia el esquema de almacenamiento definido.

Entrando en detalle en la tarea de filtrado, está compuesta por dos actividades, las cuales son: filtrado de detección y filtrado de corrección de fallas. En la primera se reciben los datos originales en una estructura estándar, se examinan y detectan posibles errores. En la segunda se aplica el proceso de corrección correspondiente. El filtrado de detección: corresponde a la detección de los errores presentes en las mediciones de cada una de las variables; haciendo uso de los filtros y restricciones que se definen para cada variable entrante. Estos filtros y restricciones son determinados por profesionales con dominio en el ámbito de aplicación del modelo. Por su parte, el filtrado de corrección de fallas recibe los datos con los errores detectados y organizados siguiendo un estándar específico. Los

filtros de corrección son definidos también por profesionales en el área de aplicación del modelo y corresponde a acciones correctivas para las mediciones de cada variable, lograr filtros correctivos para todos los casos de inconsistencia es una tarea ardua y que requiere de un conocimiento bastante especializado, tanto del área de estudio como de los datos.

Los procesos de ETL entregan datos en buenas condiciones, consistentes y de calidad; datos que pueden ser almacenados y procesados en una fase de análisis posterior. Es importante destacar la complejidad de las tareas de ETL dentro del proceso de construcción de una solución para la gestión de datos, esta complejidad se representa tanto en el costo, como en el consumo de tiempo y recursos. Cuando se cuenta con datos reales, de considerables dimensiones y con particularidades determinadas por el dominio estudiado, el proceso de ETL debe ser diseñado cuidadosamente y por medio del trabajo conjunto entre expertos del campo ambiental y del campo de los sistemas y gestión.

4.3.3 Almacenamiento

Para realizar el almacenamiento de los datos ambientales, se tiene una ventaja, y es la posibilidad de contar con sistemas de almacenamiento híbridos, es decir, que dentro del modelo se puedan combinar tanto el almacenamiento tradicional, representado en las bases de datos relacionales, como el almacenamiento en bodegas de datos, que brindan mayor escalabilidad (habilidad para reaccionar y adaptarse sin perder calidad, o bien manejar el crecimiento continuo de trabajo de manera fluida) y esquemas que facilitan la extracción de conocimiento y el almacenamiento NoSQL, que permite incluir fuentes de datos no tradicionales, como datos de imágenes satelitales, o los reportados por radares meteorológicos.

Se retoma uno de los planteamientos encontrados en la revisión del estado del arte, donde se resaltan tanto las ventajas que presentan los DBMS tradicionales como las de los enfoques NoSQL, mostrando que es viable la propuesta de construcción de sistemas de almacenamiento híbridos, que contengan varios almacenes de datos atendiendo de esta manera las necesidades de cada conjunto o fuente de datos.

4.3.4 Análisis de datos

Se puede considerar como la capa o componente más versátil dentro del modelo. Acá afloran las diferentes técnicas, tecnologías, herramientas y métodos de análisis de datos. Partiendo también de uno de los aportes revisados en el estado del arte, en los principios para el diseño de sistemas *Big Data* que proponen (Chen & Zhang, 2014), se presenta el soporte a una variedad de métodos analíticos. De acuerdo a lo anterior, y a las diferentes técnicas de análisis de datos recopiladas en el marco teórico, se propone en el modelo específico de gestión de *Big Data* ambiental, 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.

Dada la gran variedad de posibilidades que se presentan en esta capa, la dificultad de implementación de estas radica en la complejidad de los datos que se estudian. Se requiere de exploraciones cuidadosas para poder entender las relaciones y correlaciones de los fenómenos y encontrar las técnicas que mejor se adapten al dominio y que permitan extraer información relevante. Además, la aplicación de cada análisis requiere de la manipulación de conjuntos de datos particulares. El trabajo de esta capa comprende un proceso de pruebas que pueden llegar a ser inmensurables en la medida que no se conoce a ciencia cierta lo que surgirá de una combinación de técnica y datos o de la fusión de varias técnicas.

A pesar de esto, es fascinante la exploración de los datos y la generación de conocimiento, esto es lo que da impulso al planteamiento de una solución y a la aplicación del modelo.

4.3.5 Presentación de datos

La presentación de datos es fundamental, porque es la forma de llevar al público o a los interesados los resultados no solo de la capa de análisis, sino también de la capa de almacenamiento, la cual refleja a su vez las fuentes de datos. Para el caso de los datos ambientales, específicamente los hidrometeorológicos, es importante considerar tanto la presentación de datos históricos como la presentación de datos en tiempo real.

Los datos hidrometeorológicos históricos son fundamentales para entender la variabilidad climática que caracteriza una región y periodo de tiempo (Duque Méndez, Vélez Upegui, & Orozco Alzate, 2015) y constituyen las entradas para modelos de predicción de comportamientos o de reconocimiento de patrones. Estos datos históricos pueden ser presentados por medio de aplicaciones web o de escritorio que posibiliten la generación de reportes gráficos o tablas de resumen de acuerdo a los parámetros que requieran los interesados. Por su parte, presentar los datos en tiempo real es una manera de permitir un monitoreo constante de una zona particular que tiene la cobertura de una estación o red. Los datos en tiempo real pueden ser llevados a los interesados por medio de la web.

4.4 Conclusiones del capítulo

En este capítulo se introdujo la propuesta de un modelo por capas genérico para la gestión de *Big Data* y se particularizó en un modelo específico para la administración de *Big Data* ambiental, tomando en cuenta, aportes extraídos de la revisión del estado del arte y marco teórico, y de la identificación de tecnologías asociadas a *Big Data*. Se resalta que el dominio de datos a tratar puede responder o no a las cuatro características principales de *Big Data* (gran volumen de datos, alta velocidad de generación de datos, necesidad de tratar variedad en tipos de datos y necesidad de garantizar la veracidad). Se puede pensar en la implementación de un modelo de *Big Data* al contar con al menos alguna de las cuatro necesidades identificadas. Sin embargo, es importante tener en cuenta que una solución de este tipo requiere un incremento en la complejidad y costo respecto a otras soluciones tradicionales. También se puede pensar en soluciones que combinen un modelo de tratamiento de datos tradicional con tecnologías de *Big Data*, donde por ejemplo se use un modelo relacional-dimensional para el almacenamiento de los datos históricos y una solución de *Big Data* para el tratamiento de nuevos datos provenientes de la web.

5. Validación del modelo. Caso de estudio

En este capítulo se presenta la aplicación del modelo por capas específico para la administración y análisis de *Big Data* ambiental, con el fin de validar y evaluar la viabilidad de su implementación. También, con el fin de comprobar la pertinencia del modelo para el dominio de datos que se desea abordar con el desarrollo de esta tesis.

En la Figura 5-1 se puede apreciar el esquema por componentes planteado para el caso concreto, para la validación del modelo. En este esquema se traducen las capas del modelo en un caso de estudio particular, en el cual se tratan datos ambientales, de tipo hidrometeorológico.

5.1 Implementación

Para validar el modelo propuesto se construyó una solución que contempla las capas propuestas y se adapta al caso de estudio que se presentará a continuación. Para la conexión entre la capa de almacenamiento y las fuentes de datos fue necesario la adopción de una estrategia específica de extracción, transformación y carga (ETL por sus siglas en inglés de Extraction, Transformation and Load). En esta fase de ETL se permite, además de incluir diferentes fuentes de datos, contar con la capacidad de hacer un filtrado tanto de detección como de corrección, con el ánimo de garantizar la integridad y consistencia de los datos que posteriormente serán analizados. En la capa de análisis se consideró el estudio de algunas técnicas, se aplicó *Deep learning* para la predicción de comportamientos de las variables precipitación, temperatura, humedad y presión y se hizo *clustering* utilizando *Mahout* en *Hadoop*. En la capa de consumo se considera la presentación de información en tiempo real a través de la página web del Instituto de Estudios Ambientales – IDEA y la presentación de datos históricos en la plataforma CDIAC (Centro de Datos e Indicadores Ambientales de Caldas).

5.2 Caso de estudio

Para validar la arquitectura propuesta se tomó como caso de estudio los datos hidrometeorológicos recolectados por el Instituto de Estudios Ambientales – IDEA – de la Universidad Nacional de Colombia Sede Manizales. El IDEA opera una serie de redes de monitoreo ambiental ubicadas en el departamento de Caldas (Colombia), con una mayor concentración en su ciudad capital, Manizales. El departamento de Caldas hace parte de una región con un clima altamente variable, característico de las zonas tropicales (Ocampo López & Vélez Upegui, 2015). Esta condición es relevante a la hora de gestionar el componente ambiental de una región, por lo cual su estudio se convierte en un campo de trabajo de alto interés y que requiere de soluciones que contribuyan a una mejor agilidad por parte del personal experto en el tema para realizar análisis y extraer información clave.

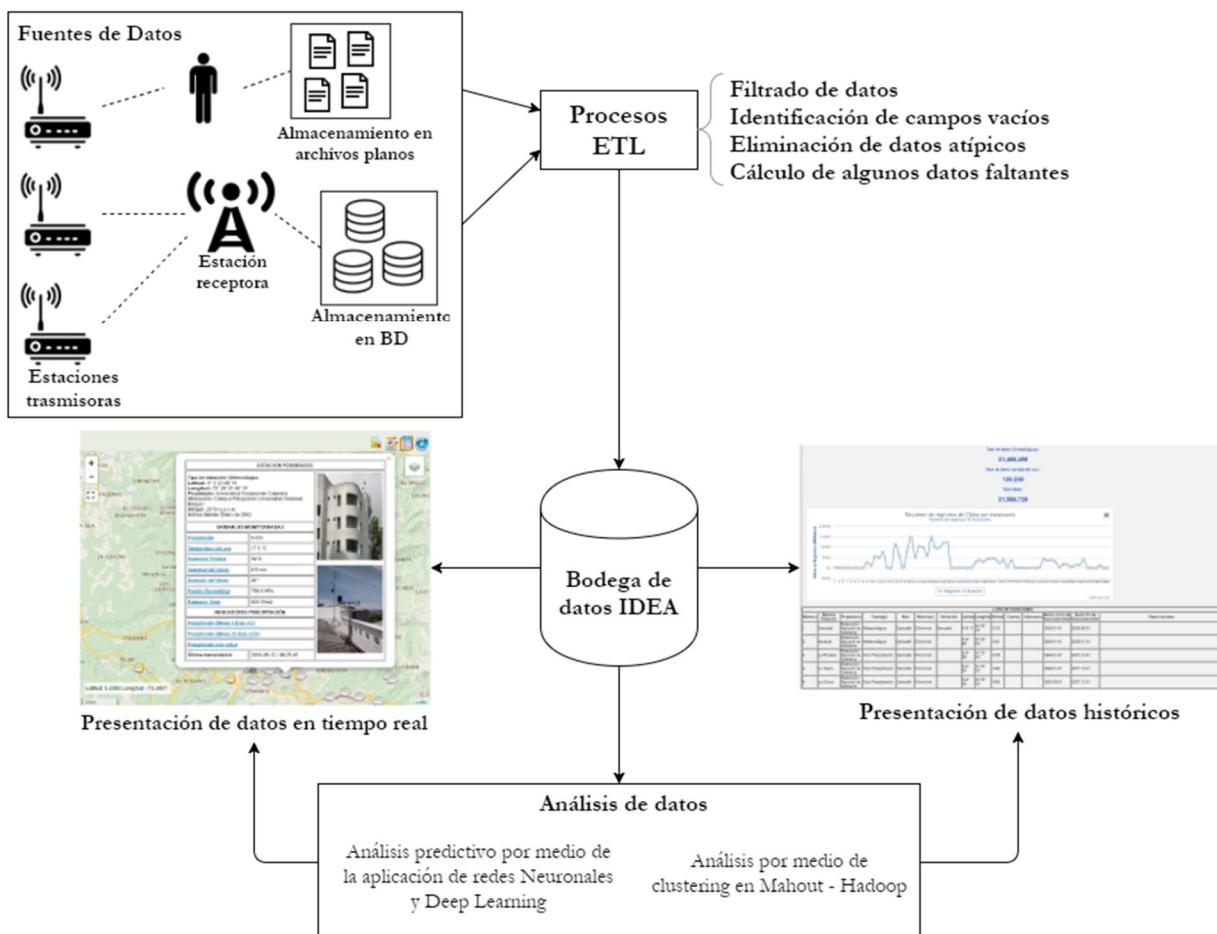


Figura 5-1: Aplicación del modelo específico al caso de estudio. Elaboración propia.

El proyecto de monitoreo hidroclimatológico lleva décadas de funcionamiento, incluso se cuenta con datos históricos de estaciones que funcionan hace más de cincuenta años. Actualmente, se cuenta con nueve redes de monitoreo principales, las cuales se sintetizan en la Tabla 5-1. En cuanto al número de estaciones, este asciende a más de ochenta, contando tanto las estaciones con datos históricos como las estaciones que actualmente trasmitten datos en tiempo real.

Tabla 5-1: Redes de monitoreo del departamento de Caldas

Red	Ubicación estaciones	Nombre de las estaciones
Red Manizales – UGR	Manizales	La Palma, Ingeominas, Alcázares, Chec-Uribe, El Carmen, Emas, Bosques del Norte, Hospital de Caldas, Quebrada San Luis Ruta 30, Quebrada Olivares – El Popal, Yarumos, Posrados, Aranjuez, Niza, Enea
SAT Cuenca Q. El Guamo	Territorio Manizales	de Quebrada El Guamo – Lavadera Los Puentes, Subestación Peralonso – CHEC, Red Sonora
SAT Cuenca Q. Olivares	Territorio Manizales	de Quebrada Olivares – Bocatoma Río Blanco, Quebrada Olivares – Bomberos Voluntarios, Alto de la Coca, El Mirador
SAT Cuenca Q. Manizales	Territorio Manizales	de Quebrada Manizales – Skinco/Colombit, Hacienda Manzanares, Finca La Paz, Quebrada Guayabal – Recinto del Pensamiento
Red hidrometeorológica cuencas Manizales	Territorio Manizales	de Quebrada Manzanares – Industrias Básicas de Caldas, Quebrada El Perro-Expoferias, Río Chinchiná – Bosque Popular El Prado, Quebrada Palogrande – Terminal de Transportes, Quebrada Marmato – Planta Intermedia CHEC, Quebrada Olivares – Aguas de Manizales, Quebrada El Triunfo – Mirador de Villapilar, Quebrada El Rosario – San Marcos de León, Quebrada La Francia – Finca Los Puentes, Quebrada Las Pavas – Autopistas del Café, Quebrada El Bohío – El Águila, Quebrada Cristales – Valles de la Alhambra
Red Caldas	Territorio departamento Caldas	de Rio Rioclaro, Rio Guacaica - Los Naranjos, Río Tapias, Camping La Palmera Risaalda, Río Supía – Los Piononos, Río Supía – Pueblo, Río Pozo, Río Pácora, Salamina – CHEC, Hogares Juveniles Campesinos Neira, Río Guacaica CHEC, Hospital de Villamaria, Quebrada Manizales Tesorito, Quebrada Olivares El Popal, Marulanda – El Páramo, Río Santo Domingo – Los Naranjos, Viejo Basurero de Manzanares, Alcaldía de Marquetalia, Río Pensilvania Microcentral, Río Doña Juana
Red Nevados	Parque Nacional Natural Los Nevados	Cisne – Santa Isabel, Molinos, Nereidas, Cumanday, Río Rioclaro

Red	Ubicación estaciones	Nombre de las estaciones
Red Udeger	Territorio departamento de Caldas	Santagueda, La Manuela, Cenicafé, El Pescador, Montevideo, El Bosque, Río Claro, El Destierro, La Batea, Jardines
Red Bocatomas CHEC	Territorio departamento de Caldas	S.BT San Francisco, S.BT La Estrella, S.BT Esmeralda, S. BT Campoalegre, S.BT Montevideo, S. BT Municipal, S.BT Sancancio, S.BT Guacaica

Las estaciones, según su tipología, transmiten datos de variables como precipitación, temperatura del aire, radiación solar, evapotranspiración, humedad relativa, presión barométrica, nivel, caudal, dirección y velocidad del viento. Las redes tienen algunas pequeñas diferencias en cuanto al proceso técnico para la transmisión de los datos, pero en general cada estación reporta mediciones, de las variables que registra, cada cinco minutos. Lo anterior puede sugerir el creciente número de datos que se tiene y la escalabilidad del problema en los próximos años, tanto por el aumento en el número de estaciones como por la inclusión de sensores para la recolección de datos de nuevas variables o la disminución del tiempo entre mediciones.

La transmisión de los datos proporciona diversas estructuras que son definidas según los requisitos de las variables medidas y el estándar establecido por la organización propietaria de cada estación. Se presentan entonces, diversas fuentes de datos, las cuales no manejan un estándar general en su estructura, haciendo necesario contar con un proceso de tratamiento previo al almacenamiento de los datos.

5.3 Aplicación de proceso de ETL

La aplicación de un proceso de ETL que se concentra en el caso particular de los datos hidrometeorológicos, proporciona la facilidad de filtrar y centralizar los datos recolectados en el departamento de Caldas. Este proceso permite garantizar el tratamiento adecuado de los datos, puesto que el esquema (ya descrito en el apartado 4.3.2) se orienta a detectar y corregir situaciones relacionadas fundamentalmente con datos faltantes, datos claramente incorrectos u outliers.

Los datos son sometidos a la fase de filtrado, contemplando tanto los filtros de detección, que permiten encontrar datos atípicos como los filtros de corrección, que permiten completar algunos datos faltantes o atípicos. Los parámetros de ambos filtros han sido

determinados según el conocimiento de expertos en temas ambientales y manejo hidrometeorológico, quienes durante años han estudiado los datos para identificar los patrones de comportamiento y rangos de mediciones válidas.

En la Figura 5-2 se presenta el modelo del proceso ETL que se aplica a los datos del caso de estudio (Duque, Hernández, Pérez, Arroyave, & Espinosa, 2016). Este proceso fue diseñado e implementado en el marco de proyectos conjuntos entre el grupo de investigación GAIA y el Grupo en Ingeniería Hidráulica y Ambiental, ambos pertenecientes al IDEA.

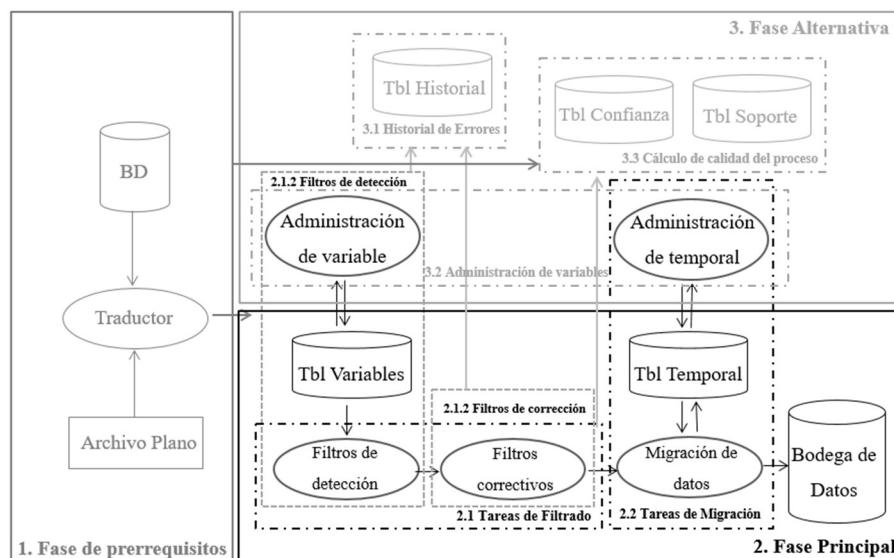


Figura 5-2: Proceso de ETL aplicado a los datos del caso de estudio. Tomado de (Duque, Hernández, Pérez, Arroyave, & Espinosa, 2016)

El proceso de ETL descrito es el paso previo al almacenamiento de los datos en la bodega de datos ambientales. A medida que ingresen nuevas variables o estaciones a la bodega es necesario y posible incluir los filtros particulares para estas, tanto detectivos como correctivos. Esta característica permite que puedan ser fácilmente integrados nuevos datos a la bodega, pero que se garantice en lo posible su consistencia y calidad; todo esto sujeto a contar con los conocimientos de los expertos en el manejo de las redes de monitoreo.

5.4 Selección de la estrategia para la capa de almacenamiento

La estrategia considerada para la capa de almacenamiento es una bodega de datos multidimensional con modelo en estrella centralizada, que fue diseñada para el almacenamiento de datos estructurados (Duque-Méndez et al., 2014) (Duque Méndez et al., 2015).

La estructura de la bodega fue concebida con el objetivo de tener un almacenamiento eficiente de los datos que garantice un tratamiento eficaz previo a la investigación meteorológica e hidrológica y que a su vez permita cubrir la oportunidad que se vislumbra en el análisis de variables hidroclimatológicas. El modelo multidimensional de estrella centralizada fue adoptado con el fin de permitir diferentes niveles de granularidad en las búsquedas que se efectúen para la extracción y futuro procesamiento de los datos. El esquema de almacenamiento consta de una tabla de hechos donde se almacena la información de las mediciones tomadas en las estaciones, y tres tablas de dimensiones, que hacen referencia a información propia de la estación, fecha y tiempo de la medición; las dimensiones mantienen una relación con la tabla de hechos a través de sus identificadores. En la Figura 5-3 se presenta la estructura de la bodega de datos ambientales.

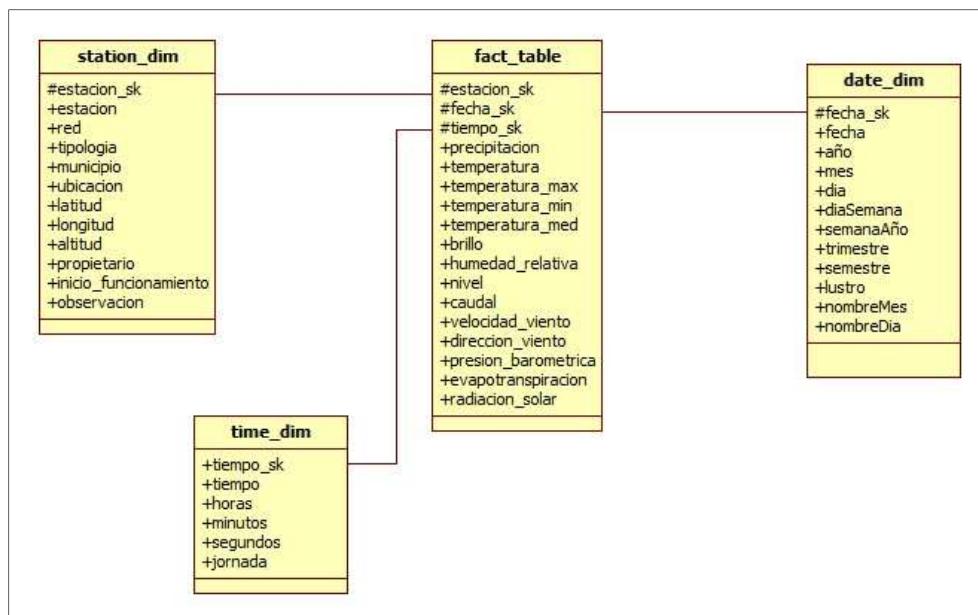


Figura 5-3: Modelo E-R de la bodega de datos ambientales. Tomado de (Duque Méndez et al., 2015)

Para la construcción de la bodega de datos se decidió utilizar el motor de bases de datos PostgreSQL, el cual fue elegido por su potencial y pensando en la gran cantidad de datos a manejar. El desarrollo de la bodega de datos ambientales lleva ya varios años de planteamiento y trabajo conjunto entre el IDEA y el grupo GAIA, contando en el momento con más de veinte millones de registros almacenados en la fact_table, y que corresponden a las mediciones de las estaciones y redes mencionadas en el caso de estudio. Estos datos constituyen una fuente valiosa de información.

Para el modelo específico de gestión de *Big Data* ambiental aplicado a los datos del caso de estudio se decidió conservar esta estrategia de almacenamiento puesto que los datos que se poseen son estructurados y pueden ser extraídos de diferentes fuentes, con lo cual la bodega de datos es una solución adecuada y que soporta el crecimiento del volumen y la inclusión de nuevas estaciones y variables, incluso la posibilidad de extender el modelo en estrella con una nueva tabla de hechos.

En el momento que se cuente con otro tipo y formato de datos se debe evaluar una nueva estrategia o tecnología que se ajuste al caso. Por ejemplo, es factible incluir almacenamiento en bases de datos llave-valor para el almacenamiento de imágenes obtenidas con satélites meteorológicos, y luego aplicar otras tecnologías *Big Data* para su análisis, recuperación y comparación.

5.5 Método para la validación de la capa de análisis

Dentro de las tareas de análisis de datos aplicados, como parte de esta capa, se incluyó las de tipo predictivo, en particular con un enfoque Deep Learning para la predicción de precipitación diaria acumulada y otras variables como temperatura, humedad y presión barométrica. También se realizaron análisis haciendo uso de tecnologías específicas de *Big Data* como *Hadoop* y *Mahout*. Con ellas se aplicaron los algoritmos de Clustering K-means y Canopy. Estos análisis se presentan a continuación.

5.5.1 Análisis predictivo usando Deep Learning

Uno de los análisis aplicados como parte de esta capa corresponde a uno de tipo predictivo que siguió un enfoque Deep Learning para la predicción de precipitación acumulada para el día siguiente a partir de datos generados en días previos.

Para esto se propuso una arquitectura compuesta por dos redes: un Autoencoder y un perceptrón multicapa (Hernández, Sánchez-Anguix, Julian, Palanca, & Duque, 2016). Los autoencoders son una técnica de Deep learning que ha sido estudiada principalmente para el análisis de imágenes, pero que se muestra prometedora para el tratamiento de series de tiempo. En la arquitectura el autoencoder tiene a cargo la tarea de selección de características en la serie de tiempo. El perceptrón multicapa es el responsable de la clasificación, tarea de predicción. Ahora se detallará cada una de las redes.

Una de las herramientas usadas habitualmente para implementar Deep learning son los autocodificadores, los cuales normalmente se implementan como una red de neuronas de tres capas, solo una capa oculta. Un autocodificador puede aprender a producir una salida exactamente igual a la información que recibe como entrada, por ello la capa de entrada y salida siempre tienen el mismo número de neuronas (Vincent, Larochelle, Bengio, & Manzagol, 2008). Lo anterior hace pensar que sería inútil su uso, pero la clave está en la capa oculta. La capa oculta brinda una representación intermedia de la información usando menos neuronas. En otras palabras, la capa oculta ofrece una versión comprimida de la información. En la arquitectura propuesta, el tipo de Autoencoder usado fue un “denoising autoencoder”.

Para la evaluación del modelo de predicción, se extrajo, a partir de la bodega de datos ambientales un *dataset* correspondiente a una serie de tiempo diaria que comprende los años entre 2002 y 2013 para la estación Posgrados, ubicada en la ciudad de Manizales. El conjunto de datos utilizado para la validación contiene un total de cuarenta y siete atributos explicativos, que incluyen mediciones de las variables temperatura, humedad relativa, presión barométrica, radiación solar, velocidad y dirección del viento, atributos de la dimensión tiempo y otros derivados de las variables principales. Este *dataset* fue dividido en entrenamiento, validación y prueba con un porcentaje de 70, 15 y 15 por ciento respectivamente. La Tabla 5-2 recoge la información detallada de dicho conjunto de datos.

Tabla 5-2: Información del conjunto de datos para el análisis predictivo. Elaboración propia

Columna de entrada	Descripción	Observación
1-3	Precipitación en los 3 últimos días atrás	Unidad de medida mm
4	Precipitación promedio de los 5 días atrás	
5	Diferencia en la precipitación registrada a las 4:00 y 24:00 horas	
6-8	Temperatura del aire en los 3 últimos días atrás	Unidad de medida °C
9	Temperatura del aire promedio de los 5 días atrás	
10	Diferencia en la temperatura registrada a las 4:00 y 24:00 horas	
11-13	Presión barométrica en los 3 últimos días atrás	Unidad de medida hPa
14	Presión barométrica promedio de los 5 días atrás	
15	Diferencia en la presión barométrica registrada a las 4:00 y 24:00 horas	
16-18	Humedad relativa en los 3 últimos días atrás	Unidad de medida %
19	Humedad relativa promedio de los 5 días atrás	
20-22	Velocidad del viento en los 3 últimos días atrás	Unidad de medida m/s
23-25	Radiación solar en los 3 últimos días atrás	Unidad de medida W/m ²
26	Radiación solar promedio de los 5 días atrás	
27	Punto de rocío	Unidad de medida °C
28-39	Valor del mes del registro de entrada	Se usan 12 componentes, por ejemplo: 010000000000 = Febrero 000000000010 = Noviembre
40-47	Dirección del viento predominante día anterior (rosa de los vientos)	N – Norte: 10000000 NE – Noreste: 01000000 E – Este: 00100000 SE – Sureste: 00010000 S – Sur: 00001000 SO – Suroeste: 00000100 O – Oeste: 00000010 NO – Noroeste: 00000001

Una vez construido el conjunto de datos, este es sometido a un proceso de normalización con el objetivo de resolver los problemas de integración de información heterogénea (Cloquell, Santamarina, & Hospitaler, 2001), es decir, llevar los datos correspondientes a diferentes variables y unidades de medida a una escala común. Los datos se sometieron a una normalización con cambio de magnitud a escala fija, siendo llevados al intervalo [0, 1] por medio del procedimiento descrito por la (Ecuación 5-1, donde V_i es la variable normalizada y a_i es la variable original.

$$v_i = \frac{a_i - \min a_i}{\max a_i - \min a_i} \quad (\text{Ecuación 5-1})$$

La Figura 5-4 presenta el flujo del método propuesto en general. Un conjunto de datos es extraído a partir de la consulta y generación de atributos desde la bodega de datos ambientales, estos datos han sido previamente recolectados por sensores ubicados en las estaciones de monitoreo ambiental y filtrados antes de poblar la bodega. Por medio de un autocodificador se hace la reducción de atributos y se genera una entrada para el perceptrón multicapa que se encarga de la tarea de predicción.

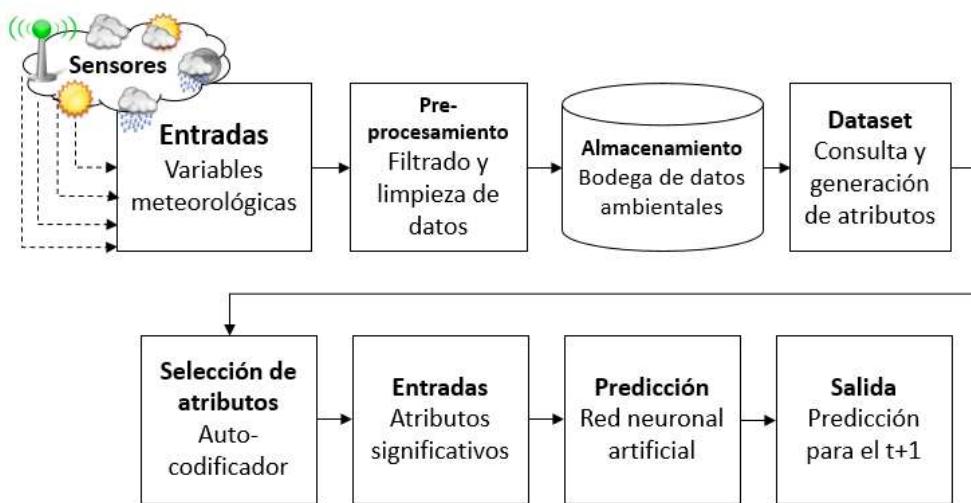


Figura 5-4: Flujo del modelo de análisis predictivo propuesto. (Hernández et al., 2016)

La Figura 5-5 presenta la arquitectura detallando las entradas y salidas y la forma como se conecta el autoencoder con el perceptrón multicapa. Un autoencoder toma una entrada $x \in [0,1]^d$ y en primer lugar la mapea, haciendo uso de un codificador, a una representación oculta de la forma $y \in [0,1]^{d'}$ por medio de un mapeo determinista, por ejemplo $y = s(Wx + b)$. Para la construcción del autoencoder se usó Theano que es una librería de Python diseñada para facilitar la escritura de modelos de Deep Learning y que da la opción de entrenarlos sobre la GPU (LISA lab., 2015). El perceptrón multicapa toma la capa intermedia del autoencoder, que incluye una representación de las variables que ingresaron inicialmente al autoencoder, ya que el autoencoder además de mapear las entradas también se encarga de reducir la dimensionalidad de los datos, de manera que toma las variables más

representativas del conjunto de datos y las transforma para entregar un *dataset* más representativo el cual facilita la tarea del perceptrón multicapa a la hora de buscar los patrones de entrenamiento para generar las predicciones con base en los datos ingresados. Por su parte, el perceptrón multicapa toma el conjunto de variables arrojadas por el autoencoder y una variable de salida que representa la variable objetivo de la fase de predicción, en este caso la precipitación; esta pareja de datos representa el patrón de entrenamiento de la red, a partir del cual se realiza la predicción de la variable de salida.

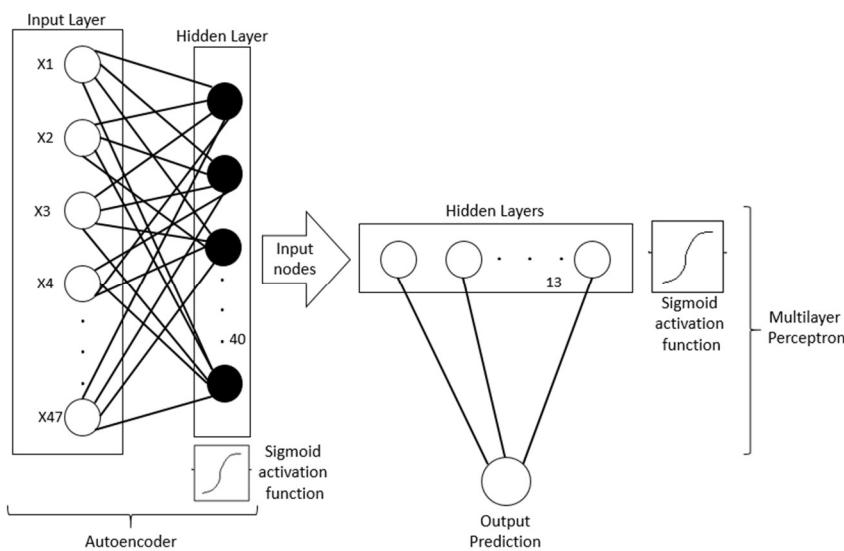


Figura 5-5: Arquitectura: Autoencoder y Perceptrón Multicapa. (Hernández et al., 2016)

Se ejecutaron múltiples entrenamientos con diferentes configuraciones para los parámetros del autoencoder y del perceptrón multicapa. Las pruebas fueron efectuadas sobre un clúster para procesamiento paralelo masivo compuesto por 72 nodos. Para las pruebas de la arquitectura, inicialmente se seleccionaron unos rangos de valores para los parámetros configurables. Una vez efectuadas las pruebas se encontró que para el caso de estudio particular la configuración con mejores resultados para el autoencoder es 40-0.9-0.1-5000 correspondiendo estos a neuronas ocultas, taza de aprendizaje, nivel de corrupción e iteraciones. Para el perceptrón multicapa la configuración que mejor se ajustó fue 13-0.3-0.9-1000 correspondiente estos valores a capas ocultas, taza de aprendizaje, momentum factor e iteraciones respectivamente. En la Tabla 5-3 se muestran todos los valores probados en cada uno de los parámetros del Autoencoder y del perceptrón.

Tabla 5-3: Valores probados para cada parámetro. Elaboración propia

	Parámetro	Valores
Autoencoder	Neuronas oculta	[11,13,19,23,29,35,40,45]
	Taza de aprendizaje	[0.1, 0.3, 0.9]
	Nivel de corrupción	[0.1, 0.3, 0.9]
	Iteraciones	[33,100,300,1000,3000,5000,9000]
Perceptrón multicapa	Capas ocultas	[10,13,19]
	Taza de aprendizaje	[0.1, 0.3, 0.9]
	Momentum factor	[0.1, 0.3, 0.9]
	Iteraciones	[33,100,300,1000,3000,5000,9000]

Para evaluar el rendimiento de la propuesta se utilizaron como criterios de medida de error, el error medio cuadrático (MSE) y la raíz del error medio cuadrático (RMSE), considerando que Y'_i es un vector de n predicciones e Y_i es el vector de valores observados correspondiente a la salida esperada de la función que genera la predicción, luego el MSE y el RMSE de la predicción pueden ser calculados como se muestra en la ecuación que sigue,

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y'_i - Y_i)^2$$

$$RMSE = \sqrt{MSE}$$

Los resultados obtenidos en la predicción por medio de la propuesta fueron comparados con predicciones obtenidas por medio de métodos ingenuos, el método Ingenuo 1 consiste en tomar como predicción la información disponible más reciente en relación al valor real y el método Ingenuo 2 consiste en tomar la media de todos los datos como el pronóstico. También se comparó con predicciones obtenidas con dos modelos revisados en el estado del arte y que fueron replicados; estos son los presentados en (Beltrán-Castro et al., 2013) y (Abhishek et al., 2012).

El modelo de (Beltrán-Castro et al., 2013) consiste en el uso de descomposición modal empírica y una Red Neuronal Feed-forward para predecir la precipitación acumulada para

el día siguiente. En (Abhishek et al., 2012) se prueban tres tipos de redes neuronales: Red Back Propagation (BP), Red de Capa Recurrente (LR - Layer Recurrent) y Red de Propagación en Cascada (CBP – Cascaded Back-Propagation). Las dos réplicas de experimentos fueron realizadas con parámetros lo más similares posibles a los usados por los autores y tomando los *dataset* de acuerdo a los parámetros presentados en los estudios.

En la Tabla 5-4 se muestra un comparativo de los resultados obtenidos para cada experimento en términos de MSE y RMSE. Las medidas de error fueron calculadas sobre los datos desnormalizados para mostrarlas en las unidades de medida originales.

Tabla 5-4: Resultados de la predicción para cada método. Elaboración propia

Método	MSE	RMSE
Autoencoder y Perceptrón Multicapa	40.11	6.33
Naive 1	132.82	11.52
Naive 2	88.43	9.40
Abhishek et al. (BP)	93.51	9.67
Abhishek et al. (LR)	81.72	9.04
Abhishek et al. (CBP)	81.36	9.02
Beltran et al.	98.73	9.94

El esquema propuesto con el autoencoder y el perceptrón multicapa superó a los modelos replicados mostrando una diferencia considerable. Los resultados de la propuesta también fueron contrastados con los dos métodos de pronóstico ingenuos superándolos sin duda.

Se hizo también un análisis predictivo para otras variables, temperatura del aire, humedad relativa y presión barométrica, siguiendo la misma arquitectura planteada para la predicción de la precipitación. Los resultados obtenidos en la predicción de estas variables son presentados en la Tabla 5-5

Tabla 5-5: Resultados para las predicciones de Temperatura, Humedad y Presión.
Elaboración propia

Temperatura del aire (°C)		
	MSE	RMSE
Autoencoder - Red	0,03	0,18
Ingenuo 1	1,35	1,16
Ingenuo 2	1,66	1,29
Humedad Relativa (%)		
	MSE	RMSE
Autoencoder - Red	21,53	4,64
Ingenuo 1	41,89	6,47
Ingenuo 2	79,61	8,92
Presión barométrica (Hpa)		
	MSE	RMSE
Autoencoder - Red	0,28	0,53
Ingenuo 1	0,19	0,44
Ingenuo 2	6,11	2,47

Con este análisis se comprobó que las arquitecturas profundas, específicamente, el uso de autotencoders y perceptrón multicapa, son una alternativa para la solución de problemas de predicción de variables meteorológicas, como precipitación, temperatura, humedad y presión barométrica, a partir del tratamiento de series de tiempo históricas.

5.5.2 Análisis de clustering con *Mahout* y *Hadoop*

También se realizó un análisis de agrupamiento haciendo uso de tecnologías *Big Data* como lo son *Hadoop* y *Mahout*. Para ello se configuró un ambiente *Hadoop* de nodo simple sobre un servidor con sistema operativo Linux. La versión de *Hadoop* utilizada fue la 2.7.2. Se configuraron también las librerías de *Mahout* junto con el ambiente *Hadoop*. La versión de *Mahout* utilizada fue la 0.12.1. Para ejecutar los algoritmos de machine learning disponibles en *Mahout* es recomendable usar *Hadoop* ya que permite paralelizar el acceso a los datos que conforman los archivos de entrenamiento y prueba a través de *MapReduce*.

Se configuraron y corrieron en *Mahout* los algoritmos de Clustering “K-means” y “Canopy”. K-means o K-medias es considerado un algoritmo de agrupamiento que tiene por objetivo la partición de un conjunto de datos de n observaciones en k grupos o clústeres, donde cada observación se lleva al clúster más cercano a la media.

Para la ejecución del análisis de Clustering, tanto con el algoritmo K-means como con el algoritmo Canopy se utilizó un *dataset* extraído de la bodega de datos ambientales, este *dataset* fue construido para las catorce estaciones pertenecientes a la Red Manizales (La Palma, Ingeominas, Alcázares, Chec-Uribe, El Carmen, Emas, Bosques del Norte, Hospital de Caldas, Quebrada San Luis Ruta 30, Yarumos, Posrados, Aranjuez, Niza, Enea). En el *dataset* se incluyó información de la ubicación de cada estación dada en términos de latitud y longitud y su respectiva altitud, también se utilizó los acumulados diarios de precipitación y la temperatura diaria promedio.

Para el algoritmo de K-means se parametrizaron los siguientes valores:

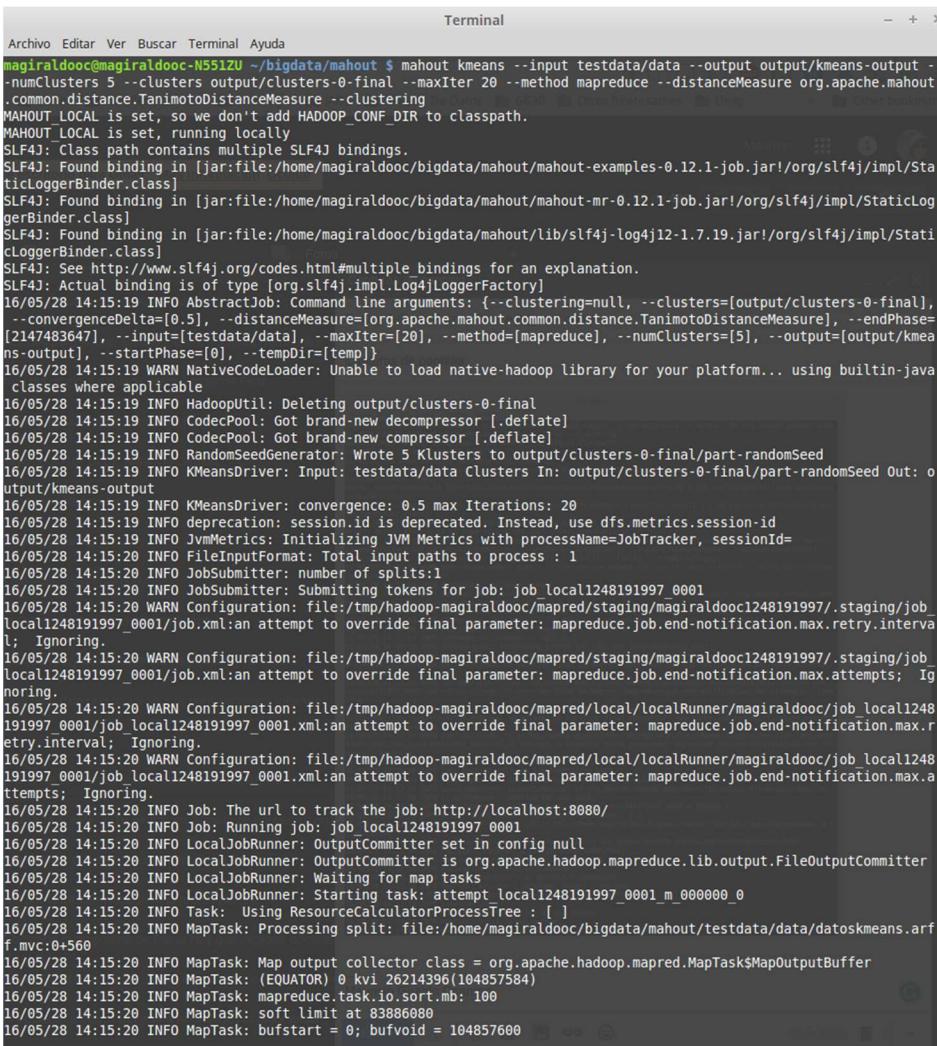
- Distancia Tanimoto
- Iteraciones 20
- Método Map Reduce

Se hizo un agrupamiento inicial con un k=4. Los resultados arrojados por *Mahout* y todo el proceso de ejecución del algoritmo puede ser visualizados en la consola, como se muestra en la Figura 5-7.

Los resultados también son exportados en archivos con formato .csv, .txt ó .graphml. En la Figura 5-6 se presentan los resultados obtenidos y exportados en un archivo de texto plano.

```
[{"identifier": "VL-4", "r": [0.471], "c": [4.667], "n": 3}
  Weight : [props - optional]: Point:
  1.0 : [distance=0.023255813953488635]: [4.0]
  1.0 : [distance=0.004739336492891155]: [5.0]
 {"identifier": "VL-6", "r": [0.471], "c": [6.667], "n": 3}
  Weight : [props - optional]: Point:
  1.0 : [distance=0.010989010989011283]: [6.0]
  1.0 : [distance=0.002375296912114133]: [7.0]
  1.0 : [distance=0.032258064516129115]: [8.0]
 {"identifier": "VL-2", "r": [0.829], "c": [2.25], "n": 4}
  Weight : [props - optional]: Point:
  1.0 : [distance=0.4098360655737705]: [1.0]
  1.0 : [distance=0.013698630136986356]: [2.0]
  1.0 : [distance=0.07692307692307687]: [3.0]
 {"identifier": "VL-7", "r": [2.118], "c": [10.625], "n": 8}
  Weight : [props - optional]: Point:
  1.0 : [distance=0.02687231674352042]: [9.0]
  1.0 : [distance=0.00366300366300365]: [10.0]
  1.0 : [distance=0.0012017625851248326]: [11.0]
  1.0 : [distance=0.014611761864509076]: [12.0]
  1.0 : [distance=0.039234865775459205]: [13.0]
  1.0 : [distance=0.0711288906234755]: [14.0]
```

Figura 5-6: Resultados K-means con *Mahout* exportados a archivo plano



```

Archivo Editar Ver Buscar Terminal Ayuda
magiraldooc@magiraldooc-N551ZU ~ /bigdata/mahout $ mahout kmeans --input testdata/data --output output/kmeans-output \
--numClusters 5 --clusters output/clusters-0-final --maxIter 20 --method mapreduce --distanceMeasure org.apache.mahout.common.distance.TanimotoDistanceMeasure --clustering
MAHOUT_LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT_LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/magiraldooc/bigdata/mahout-examples-0.12.1-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/magiraldooc/bigdata/mahout-mr-0.12.1-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/magiraldooc/bigdata/mahout/lib/slf4j-log4j12-1.7.19.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/05/28 14:15:19 INFO AbstractJob: Command line arguments: {-clustering=null, --clusters=[output/clusters-0-final], \
--convergenceDelta=[0.5], --distanceMeasure=[org.apache.mahout.common.distance.TanimotoDistanceMeasure], --endPhase=[2147483647], --input=[testdata/data], --maxIter=[20], --method=[mapreduce], --numClusters=[5], --output=[output/kmeans-output], --startPhase=[0], --tempDir=[temp]}
16/05/28 14:15:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/05/28 14:15:19 INFO HadoopUtil: Deleting output/clusters-0-final
16/05/28 14:15:19 INFO CodecPool: Got brand-new decompressor [.deflate]
16/05/28 14:15:19 INFO CodecPool: Got brand-new compressor [.deflate]
16/05/28 14:15:19 INFO RandomSeedGenerator: Wrote 5 Klusters to output/clusters-0-final/part-randomSeed
16/05/28 14:15:19 INFO KMeansDriver: Input: testdata/data Clusters In: output/clusters-0-final/part-randomSeed Out: o
utput/kmeans-output
16/05/28 14:15:19 INFO KMeansDriver: convergence: 0.5 max Iterations: 20
16/05/28 14:15:19 INFO deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/05/28 14:15:19 INFO JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/05/28 14:15:20 INFO FileInputFormat: Total input paths to process : 1
16/05/28 14:15:20 INFO JobSubmitter: number of splits:1
16/05/28 14:15:20 INFO JobSubmitter: Submitting tokens for job: job_local1248191997_0001
16/05/28 14:15:20 WARN Configuration: file:/tmp/hadoop-magiraldooc/mapred/staging/magiraldooc1248191997/.staging/job_local1248191997_0001/job.xml:an attempt to override final parameter: mapreduce.job.end-notification.max.retry.interval; Ignoring.
16/05/28 14:15:20 WARN Configuration: file:/tmp/hadoop-magiraldooc/mapred/staging/magiraldooc1248191997/.staging/job_local1248191997_0001/job.xml:an attempt to override final parameter: mapreduce.job.end-notification.max.attempts; Ignoring.
16/05/28 14:15:20 WARN Configuration: file:/tmp/hadoop-magiraldooc/mapred/local/localRunner/magiraldooc/job_local1248191997_0001/job_local1248191997_0001.xml:an attempt to override final parameter: mapreduce.job.end-notification.max.retry.interval; Ignoring.
16/05/28 14:15:20 WARN Configuration: file:/tmp/hadoop-magiraldooc/mapred/local/localRunner/magiraldooc/job_local1248191997_0001/job_local1248191997_0001.xml:an attempt to override final parameter: mapreduce.job.end-notification.max.attempts; Ignoring.
16/05/28 14:15:20 INFO Job: The url to track the job: http://localhost:8080/
16/05/28 14:15:20 INFO Job: Running job: job_local1248191997_0001
16/05/28 14:15:20 INFO LocalJobRunner: OutputCommitter set in config null
16/05/28 14:15:20 INFO LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
16/05/28 14:15:20 INFO LocalJobRunner: Waiting for map tasks
16/05/28 14:15:20 INFO LocalJobRunner: Starting task: attempt_local1248191997_0001_m_000000_0
16/05/28 14:15:20 INFO Task: Using ResourceCalculatorProcessTree : []
16/05/28 14:15:20 INFO MapTask: Processing split: file:/home/magiraldooc/bigdata/mahout/testdata/data/datoskmeans.arff.mvc:0+566
16/05/28 14:15:20 INFO MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/05/28 14:15:20 INFO MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/05/28 14:15:20 INFO MapTask: mapreduce.task.io.sort.mb: 100
16/05/28 14:15:20 INFO MapTask: soft limit at 83886080
16/05/28 14:15:20 INFO MapTask: bufstart = 0; bufvoid = 104857600

```

Figura 5-7: Resultados en consola para el algoritmo K-means con *Mahout*

Sin embargo, como la lectura de estos archivos no es muy cómoda, se han traducido estos resultados a una forma de visualización gráfica donde se muestra la ubicación concreta de las estaciones de la Red Manizales sobre el mapa y se han marcado en un color diferente cada uno de los clúster formados; esto se aprecia en la Figura 5-8.



Figura 5-8: Clústeres formados con un $k=4$. Elaboración propia

Para el segundo acercamiento se utilizó un $k=5$. En la Figura 5-9 y Figura 5-10 se presentan los resultados obtenidos, por medio de una captura del archivo de texto plano y en la imagen construida en el mapa a partir de estos.

```
{
  "identifier": "VL-11", "r": [0.829], "c": [12.75], "n": 4
    Weight : [props - optional]: Point:
      1.0 : [distance=0.00366300366300365]: [12.0]
      1.0 : [distance=3.7693177534869893E-4]: [13.0]
      1.0 : [distance=0.008677542519958337]: [14.0]
  {"identifier": "VL-10", "r": [0.471], "c": [10.667], "n": 3}
    Weight : [props - optional]: Point:
      1.0 : [distance=0.004149377593361092]: [10.0]
      1.0 : [distance=9.460737937558861E-4]: [11.0]
  {"identifier": "VL-2", "r": [1.02], "c": [3.4], "n": 5}
    Weight : [props - optional]: Point:
      1.0 : [distance=0.22374429223744274]: [2.0]
      1.0 : [distance=0.015444015444015413]: [3.0]
      1.0 : [distance=0.025787965616045794]: [4.0]
      1.0 : [distance=0.13087934560327208]: [5.0]
  {"identifier": "VL-8", "r": [1.166], "c": [7.8], "n": 5}
    Weight : [props - optional]: Point:
      1.0 : [distance=0.06474820143884907]: [6.0]
      1.0 : [distance=0.011585807385952274]: [7.0]
      1.0 : [distance=6.406149903909197E-4]: [8.0]
      1.0 : [distance=0.02010050251256279]: [9.0]
  {"identifier": "VL-0", "r": [], "c": [1.0], "n": 2}
    Weight : [props - optional]: Point:
      1.0 : [distance=0.0]: [1.0]
}
```

Figura 5-9: Resultados K-means con *Mahout* exportados a archivo plano para un $k=5$

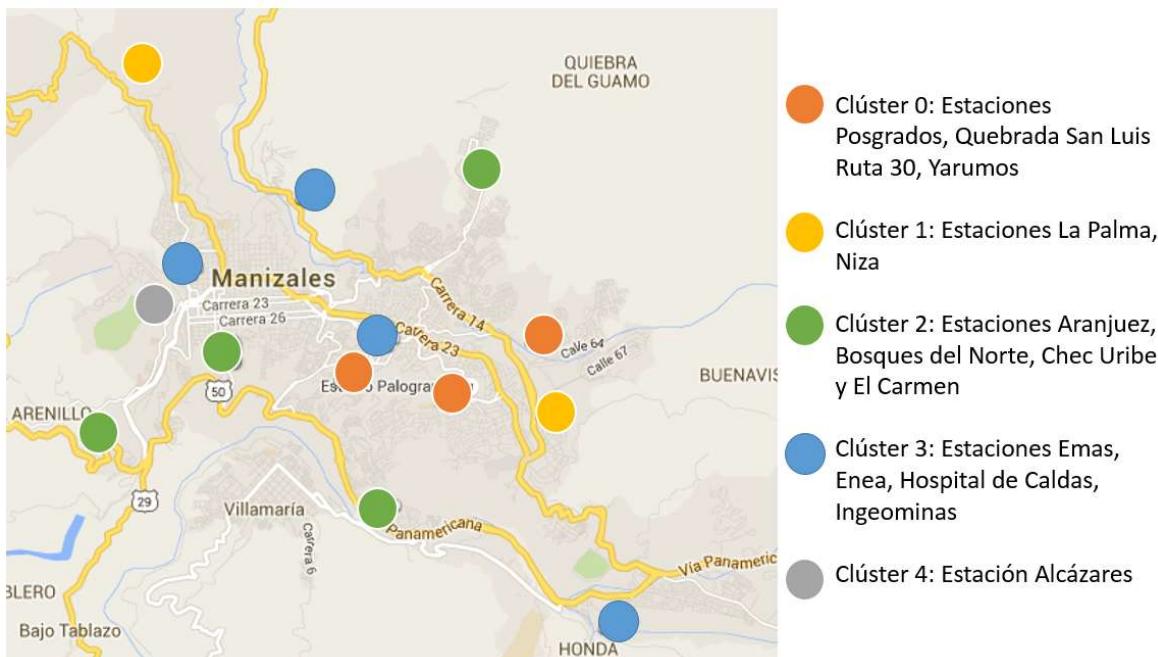
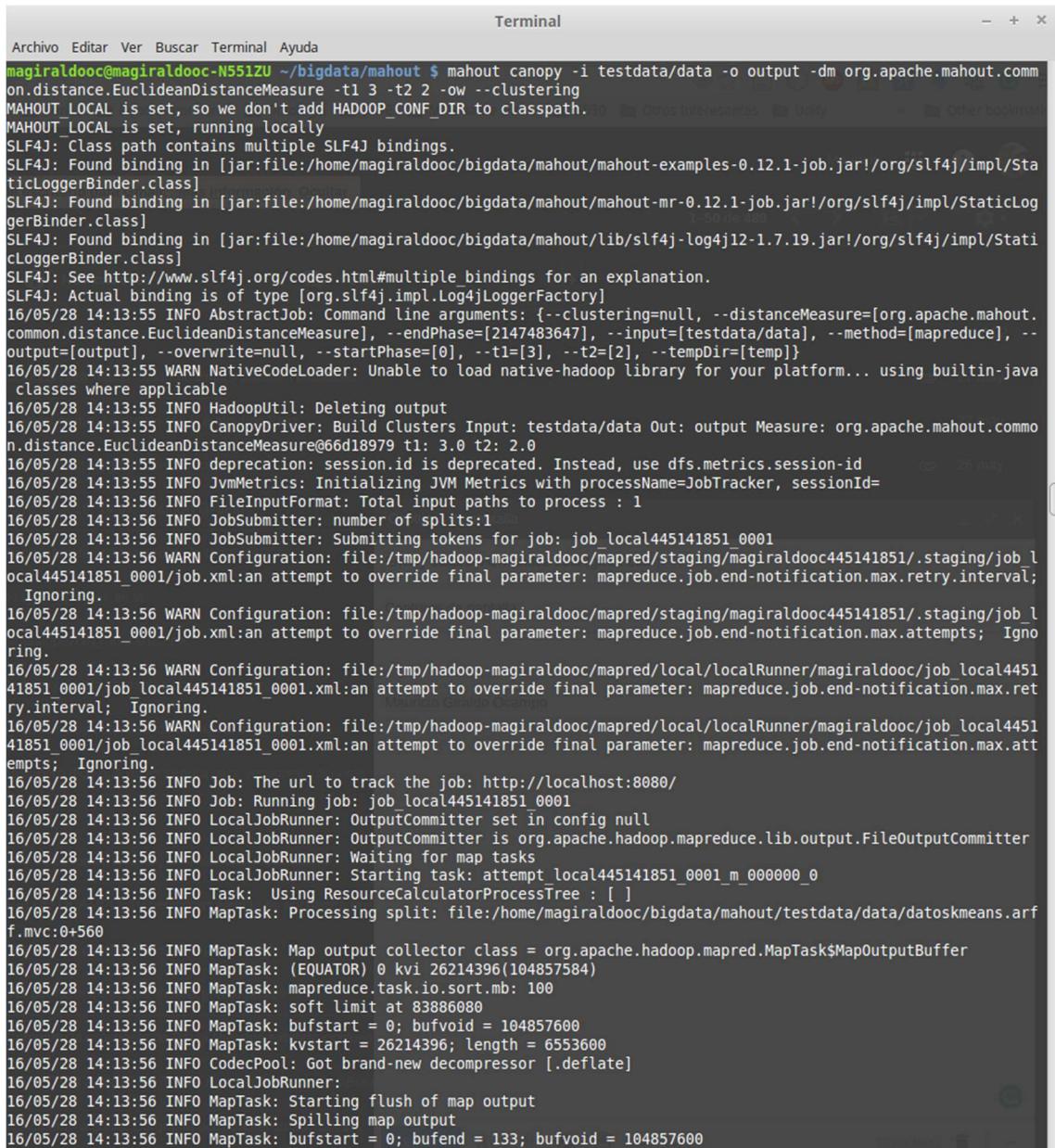


Figura 5-10: Clústeres formados con un $k=5$. Elaboración propia

Por su parte, Canopy, es también un algoritmo de agrupamiento o Clustering, pero su principal característica radica en la posibilidad que brinda de reducir considerablemente el número de cálculos necesarios en otros algoritmos como por ejemplo el k-means, esta reducción la logra al introducir un proceso previo de generación de clúster superpuestos o *canopies* a partir de una métrica de cálculo sencillo. Para este algoritmo se usó una distancia euclidiana, en este caso no se define el número de clústeres, sino que estos son determinados por el algoritmo. Las iteraciones fueron 20 en total.

En la Figura 5-11 se presentan el proceso de ejecución del algoritmo en la consola. Mientras que en la Figura 5-12 se presentan los resultados exportados en un archivo de texto plano y también son presentados de forma gráfica en la Figura 5-13. Se obtuvieron un total de seis clúster. El cluster 0 comprende las estaciones Niza, Posgrados, Quebrada San Luis Ruta 30, Yarumos, el clúster 1 las estaciones Hospital de Caldas, Ingeominas, La Palma, el clúster 2 las estaciones Emas, Enea, en el clúster 3 las estaciones Chec Uribe, El Carmen, en el clúster 4 la estación Bosques del Norte y en el clúster 5 las estaciones Alcázares, Aranjuez.

Capítulo 5



The screenshot shows a terminal window titled "Terminal" with the following command and its output:

```
Archivo Editar Ver Buscar Terminal Ayuda
magiraldooc@magiraldooc-N551ZU ~/bigdata/mahout $ mahout canopy -i testdata/data -o output -dm org.apache.mahout.common.distance.EuclideanDistanceMeasure -t1 3 -t2 2 -ow --clustering
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.30  Otras interesantes  Unity  Other bookmarks
MAHOUT LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/magiraldooc/bigdata/mahout/mahout-examples-0.12.1-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/magiraldooc/bigdata/mahout/mahout-mr-0.12.1-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/magiraldooc/bigdata/mahout/lib/slf4j-log4j12-1.7.19.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/05/28 14:13:55 INFO AbstractJob: Command line arguments: {--clustering=null, --distanceMeasure=[org.apache.mahout.common.distance.EuclideanDistanceMeasure], --endPhase=[2147483647], --input=[testdata/data], --method=[mapreduce], --output=[output], --overwrite=null, --startPhase=[0], --t1=[3], --t2=[2], --tempDir=[temp]}
16/05/28 14:13:55 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/05/28 14:13:55 INFO HadoopUtil: Deleting output
16/05/28 14:13:55 INFO CanopyDriver: Build Clusters Input: testdata/data Out: output Measure: org.apache.mahout.common.distance.EuclideanDistanceMeasure@66d18979 t1: 3.0 t2: 2.0
16/05/28 14:13:55 INFO deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/05/28 14:13:55 INFO JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/05/28 14:13:56 INFO FileInputFormat: Total input paths to process : 1
16/05/28 14:13:56 INFO JobSubmitter: number of splits:1
16/05/28 14:13:56 INFO JobSubmitter: Submitting tokens for job: job_local445141851_0001
16/05/28 14:13:56 WARN Configuration: file:/tmp/hadoop-magiraldooc/mapred/staging/magiraldooc445141851/.staging/job_local445141851_0001/job.xml:an attempt to override final parameter: mapreduce.job.end-notification.max.retry.interval; Ignoring.
16/05/28 14:13:56 WARN Configuration: file:/tmp/hadoop-magiraldooc/mapred/staging/magiraldooc445141851/.staging/job_local445141851_0001/job.xml:an attempt to override final parameter: mapreduce.job.end-notification.max.attempts; Ignoring.
16/05/28 14:13:56 WARN Configuration: file:/tmp/hadoop-magiraldooc/mapred/local/localRunner/magiraldooc/job_local445141851_0001/job_local445141851_0001.xml:an attempt to override final parameter: mapreduce.job.end-notification.max.retry.interval; Ignoring.
16/05/28 14:13:56 WARN Configuration: file:/tmp/hadoop-magiraldooc/mapred/local/localRunner/magiraldooc/job_local445141851_0001/job_local445141851_0001.xml:an attempt to override final parameter: mapreduce.job.end-notification.max.attempts; Ignoring.
16/05/28 14:13:56 INFO Job: The url to track the job: http://localhost:8080/
16/05/28 14:13:56 INFO Job: Running job: job_local445141851_0001
16/05/28 14:13:56 INFO LocalJobRunner: OutputCommitter set in config null
16/05/28 14:13:56 INFO LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
16/05/28 14:13:56 INFO LocalJobRunner: Waiting for map tasks
16/05/28 14:13:56 INFO LocalJobRunner: Starting task: attempt_local445141851_0001_m_000000_0
16/05/28 14:13:56 INFO Task: Using ResourceCalculatorProcessTree : []
16/05/28 14:13:56 INFO MapTask: Processing split: file:/home/magiraldooc/bigdata/mahout/testdata/data/datoskmeans.arff.mvc:0+560
16/05/28 14:13:56 INFO MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/05/28 14:13:56 INFO MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/05/28 14:13:56 INFO MapTask: mapreduce.task.io.sort.mb: 100
16/05/28 14:13:56 INFO MapTask: soft limit at 83886080
16/05/28 14:13:56 INFO MapTask: bufstart = 0; bufvoid = 104857600
16/05/28 14:13:56 INFO MapTask: kvstart = 26214396; length = 6553600
16/05/28 14:13:56 INFO CodecPool: Got brand-new decompressor [.deflate]
16/05/28 14:13:56 INFO LocalJobRunner:
16/05/28 14:13:56 INFO MapTask: Starting flush of map output
16/05/28 14:13:56 INFO MapTask: Spilling map output
16/05/28 14:13:56 INFO MapTask: bufstart = 0; bufend = 133; bufvoid = 104857600
```

Figura 5-11: Resultados en consola para el algoritmo Canopy con *Mahout*

```
{
  "identifier": "C-0", "r": [0.75], "c": [12.75], "n": 2
    Weight : [props - optional]: Point:
    1.0 : [distance=1.75]: [11.0]
    1.0 : [distance=0.75]: [12.0]
    1.0 : [distance=0.25]: [13.0]
    1.0 : [distance=1.25]: [14.0]
  {"identifier": "C-1", "r": [1.0], "c": [9.0], "n": 2}
    Weight : [props - optional]: Point:
    1.0 : [distance=1.0]: [8.0]
    1.0 : [distance=0.0]: [9.0]
    1.0 : [distance=1.0]: [10.0]
  {"identifier": "C-2", "r": [1.0], "c": [7.0], "n": 2}
    Weight : [props - optional]: Point:
    1.0 : [distance=1.0]: [6.0]
    1.0 : [distance=0.0]: [7.0]
  {"identifier": "C-3", "r": [1.0], "c": [5.0], "n": 2}
    Weight : [props - optional]: Point:
    1.0 : [distance=1.0]: [4.0]
    1.0 : [distance=0.0]: [5.0]
  {"identifier": "C-4", "r": [1.0], "c": [3.0], "n": 2}
    Weight : [props - optional]: Point:
    1.0 : [distance=0.0]: [3.0]
  {"identifier": "C-5", "r": [], "c": [2.0], "n": 1}
    Weight : [props - optional]: Point:
    1.0 : [distance=1.0]: [1.0]
    1.0 : [distance=0.0]: [2.0]
}
```

Figura 5-12: Resultados del algoritmo Canopy con *Mahout* exportados a un archivo de texto plano.

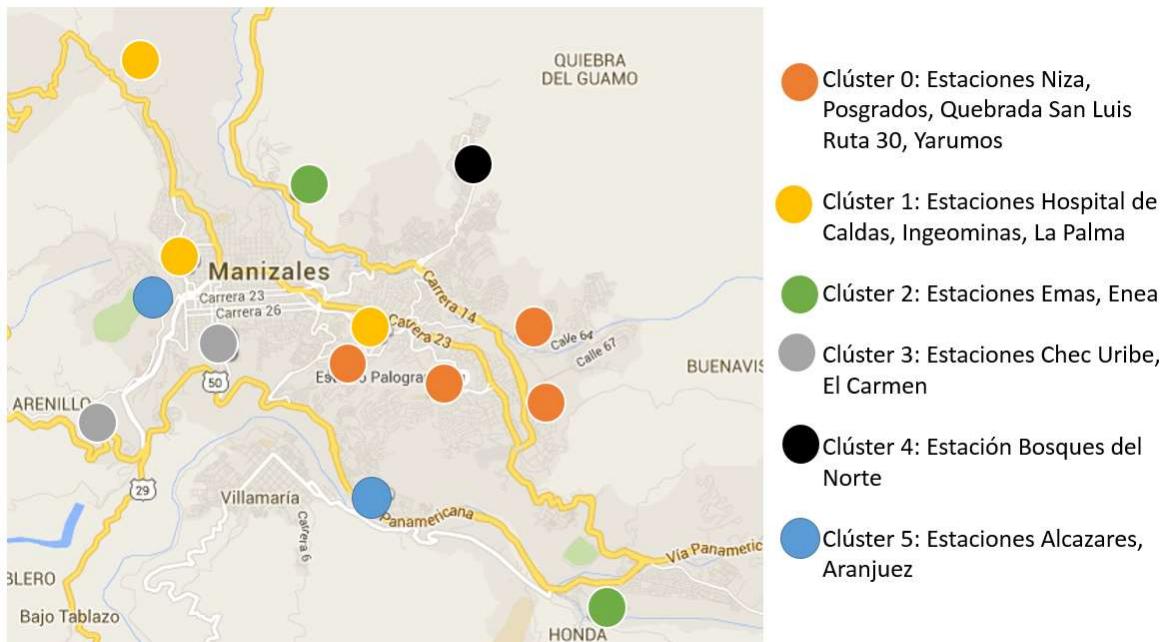


Figura 5-13: Clústeres formados con el algoritmo Canopy. Elaboración propia

Los resultados obtenidos en el análisis predictivo y de agrupamiento muestran las diferentes posibilidades que se pueden aplicar sobre los datos del caso de estudio. El número de técnicas de análisis disponibles es bastante amplio y el modelo permite que se adhieran las técnicas que se consideren necesarias o que se quieran explorar.

5.6 Presentación de resultados para la capa de consumo

Para la capa de consumo se cuenta con dos estrategias o herramientas fundamentales. Una de ellas es la presentación de datos en tiempo real a través de la página del Instituto (idea.manizales.unal.edu.co). Para ello se cuenta con dos aplicativos de estado del tiempo, uno para el departamento de Caldas y otro para la ciudad de Manizales, las redes presentadas en la Tabla 5-1 se distribuyen entre estos dos aplicativos. Los aplicativos de estado del tiempo despliegan un mapa donde se encuentran geoposicionadas las estaciones de cada una de las redes. Cada estación asume un ícono que corresponde a una representación gráfica que refleja el clima del punto donde se encuentra ubicada.

Las estaciones se encuentran distribuidas en capas, cada capa corresponde a una red; las capas se pueden habilitar o deshabilitar y de esta forma facilitar la visualización de una red o redes en particular (ver Figura 5-14).

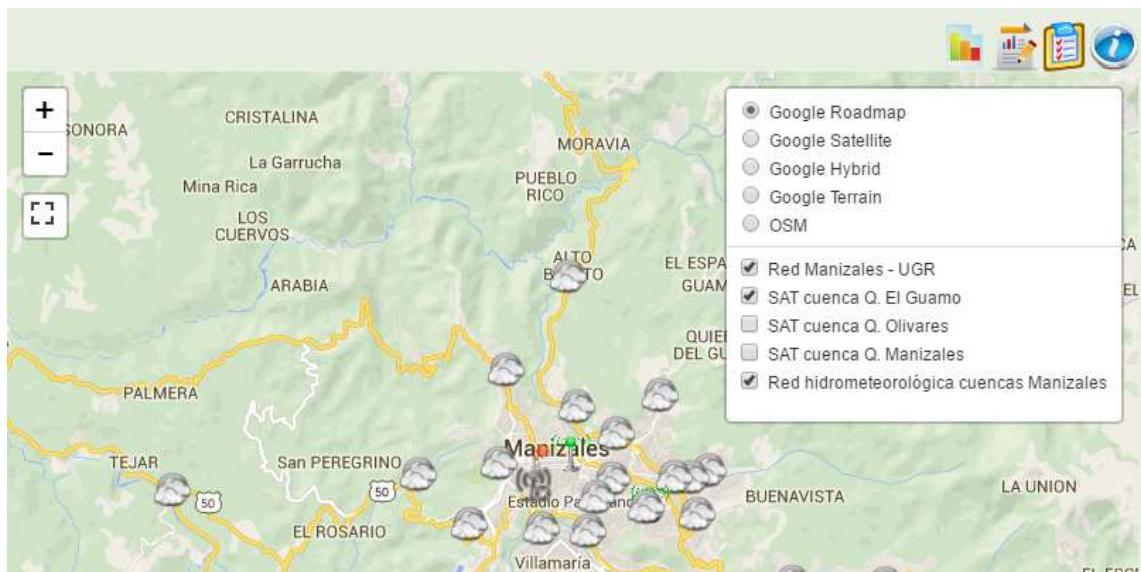


Figura 5-14: Visualización de las capas en el aplicativo de estado del tiempo. Obtenido de <http://idea.manizales.unal.edu.co/index.php/estado-tiempo-manizales>

Se puede realizar una ampliación de los datos del clima registrado en cada estación haciendo clic sobre el ícono que la representa, esta acción despliega una ventana que presenta además de los datos de la última medición registrada para cada variable, la información de la ficha técnica de la estación (tipo de estación, altitud, latitud, longitud, propietario, lugar de ubicación, inicio de funcionamiento). Un ejemplo de esto se presenta en la Figura 5-15.

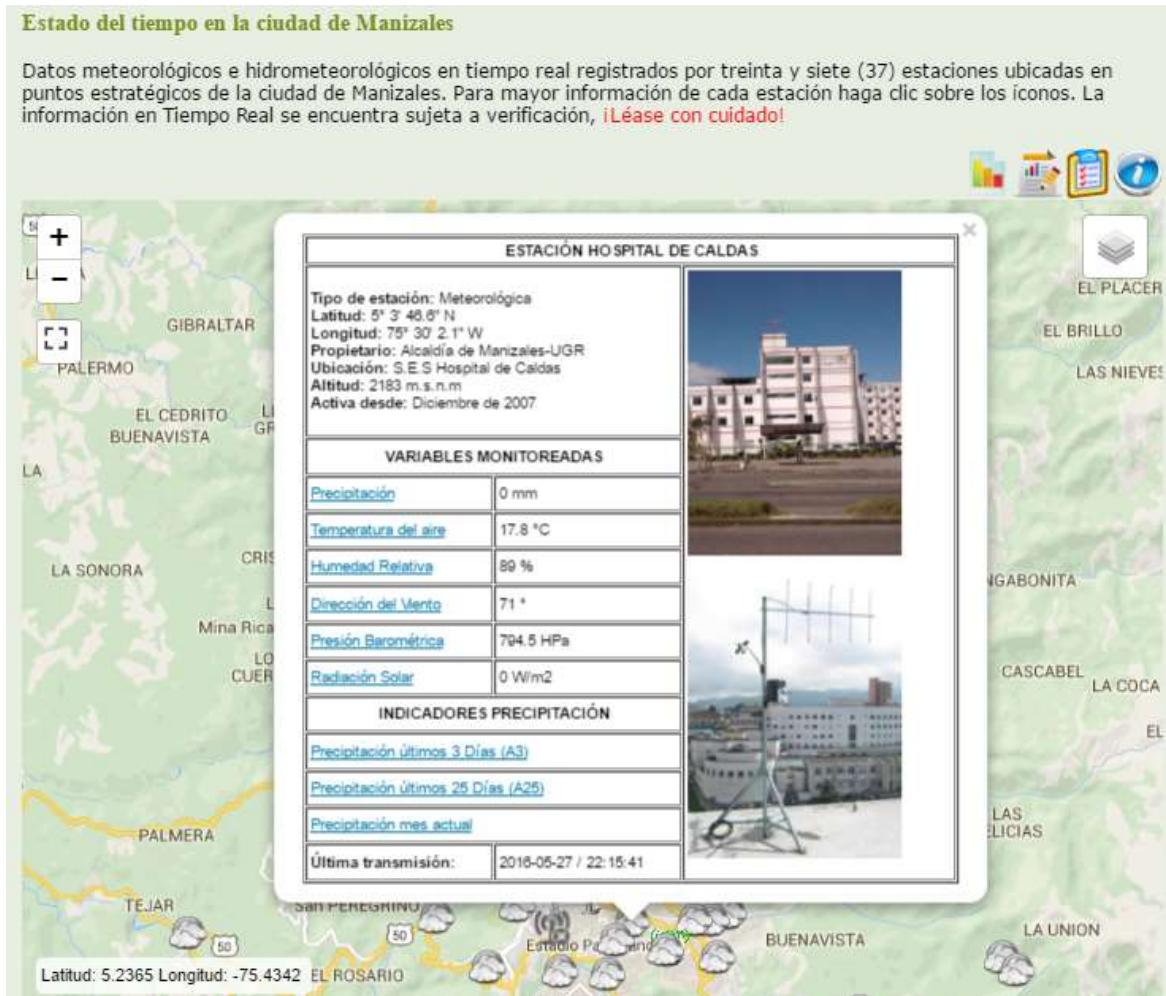


Figura 5-15: Presentación de datos detallados de cada estación en tiempo real. Obtenido de <http://idea.manizales.unal.edu.co/index.php/estado-tiempo-manizales>

Para cada variable presentada en la ventana desplegada se cuenta con la posibilidad de visualizar por medio de una gráfica el comportamiento de la variable en las últimas veinticuatro horas. Para los indicadores de precipitación también se presentan gráficas

correspondiente a los antecedentes A3 (precipitación de los tres días), A25 (precipitación de los últimos veinticinco días) y precipitación del mes en curso. Estas gráficas son interactivas, permiten realizar zoom, revisar puntos concretos de las líneas o barras graficadas y deshabilitar o habilitar algunas opciones. Además se presentan unos estadísticos básicos en la parte inferior, respecto a medias, máximas y mínimas. En la Figura 5-16 se presenta un ejemplo de una gráfica de comportamiento de las últimas 24 horas para la variable temperatura del aire y una gráfica del antecedente A25.

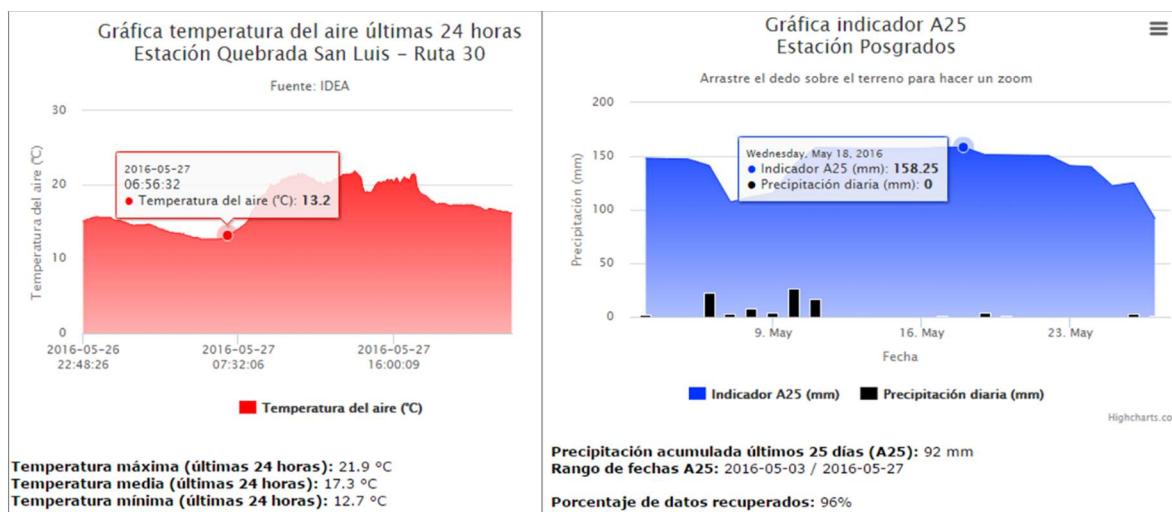


Figura 5-16: Ejemplo de gráficas para la temperatura e indicador A25. Obtenido de <http://idea.manizales.unal.edu.co/index.php/estado-tiempo-manizales>

Otra de las estrategias es la presentación de los datos históricos y de indicadores a partir de estos. En la plataforma CDIAC, la cual se encuentra disponible en la web, accediendo al enlace <http://cdiac.manizales.unal.edu.co/> se presenta un sistema de consulta de datos y un sistema de generación de indicadores hidroclimatológicos. Estas herramientas se desarrollaron en el marco del proyecto Línea base ambiental de Caldas, que se suscribió entre la Universidad Nacional de Colombia y la Corporación Autónoma Regional de Caldas, Corpocaldas. En el sistema de consulta se permite hacer peticiones de datos almacenados en la bodega de datos ambientales eligiendo la estación, variables y rango de tiempo que se quiere visualizar, además se cuenta con la posibilidad de descargar estos datos en archivos planos. El sistema de generación de indicadores brinda la posibilidad de escoger la estación, indicador, rango de fechas y dimensión tiempo por la cual se desea agrupar los datos. Los resultados de la consulta se pueden revisar tanto en tablas como de forma

gráfica. En la Figura 5-17 se presenta la vista principal del sistema de consulta de datos para un usuario autenticado.

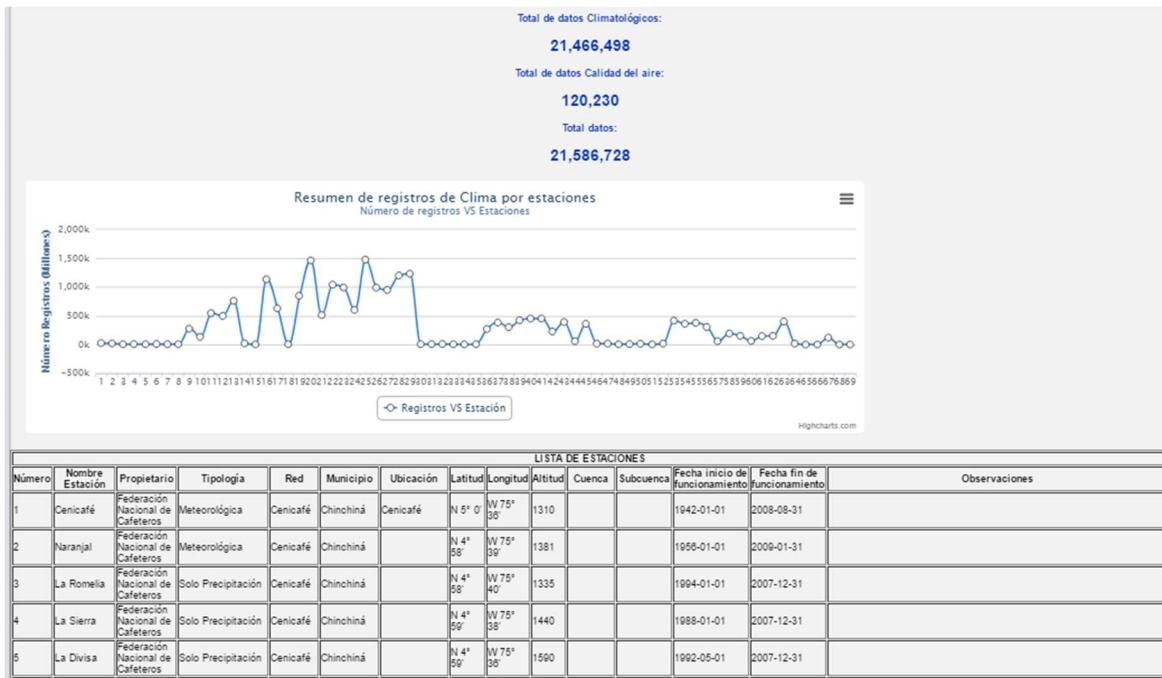


Figura 5-17: Presentación de datos históricos. Obtenido de cdiac.manizales.unal.edu.co

En esta capa también deben ser presentados los productos obtenidos en la capa de análisis después de la aplicación de diferentes técnicas; sin embargo, deben ser llevados a una interfaz que permita su fácil visualización y comprensión, una presentación que permita que los expertos en el dominio, sin ser conocedores de las técnicas aplicadas, puedan determinar la relevancia y aplicabilidad de los resultados. Existen herramientas que presentan los resultados de forma gráfica directamente, también hay otras herramientas que leen los resultados en ciertos formatos de texto plano y los llevan a un esquema gráfico.

5.7 Conclusiones del capítulo

El modelo específico aplicado en el caso de estudio demuestra que es posible implementar la propuesta para conformar un sistema de administración de datos ambientales, en este caso, hidrometeorológicos. También se hace factible combinar tecnologías tradicionales

de tratamiento de datos con tecnologías asociadas a *Big Data*. El modelo se ha probado con el esquema propuesto en un dominio de datos ambientales estructurados, pero se puede pensar en la inclusión de nuevas fuentes de datos no estructurados o semi-estructurados, como por ejemplo, datos provenientes de radares y satélites meteorológicos o de mapas de suelos, aplicando nuevas tecnologías para su almacenamiento e incluso para su análisis. Se presenta la validación de algunos algoritmos de *Mahout* en *Hadoop*, a partir de los cuales se han obtenido resultados de análisis en la aplicación al caso de estudio, también se presentan los resultados de la aplicación del enfoque Deep learning para el análisis predictivo; se generaron pronósticos para las variables precipitación, temperatura del aire, humedad relativa y presión barométrica. Estos son una fuente valiosa para futuros análisis y procesos de predicción.

6. Conclusiones y trabajos futuros

6.1 Conclusiones

Con el desarrollo de esta tesis se buscaba hacer frente a algunas limitaciones encontradas respecto a la administración de grandes volúmenes de datos ambientales (hidrometeorológicos). En particular, se logró como resultado la definición de un modelo por capas para la gestión y tratamiento de datos en el dominio ambiental. Cada una de las capas del modelo propuesto tiene elementos modulares que se pueden incluir, cambiar o ampliar. La aplicación del modelo por capas permite presentar información relevante para el análisis de datos de este campo de dominio acordes a la región y que podrán ser utilizados en la toma de decisiones o para el entendimiento de los fenómenos que están detrás de este tipo de datos.

En este trabajo se aprovechan algunas ventajas de las tecnologías asociadas a *Big Data* para el tratamiento de datos, las cuales integran paradigmas y enfoques de los cuales se obtienen resultados satisfactorios con buenos tiempos de respuesta, pero que pueden estar asociados a la necesidad de recursos de cómputo físicos de elevado alcance. En particular, en cuanto a las tecnologías asociadas al enfoque de *Big Data*, se vislumbran grandes oportunidades y retos en su utilización y adaptación a diferentes dominios de datos, por lo cual siguen siendo materia de investigación y discusión. A pesar que las aplicaciones generalmente no cubren las 4V (volumen, veracidad, variedad y velocidad), el enfoque *Big Data* ataca facetas de las nuevas características de los datos que se enfrentan permanentemente.

Teniendo en cuenta los objetivos planteados en la propuesta de tesis, estos se lograron cumplir de la siguiente manera:

Se realizó revisión del marco teórico y estado del arte que permitió identificar espacios de investigación bastante amplios, tratando de atacar con esta tesis una porción de estos, al proponer un modelo de administración y análisis de datos ambientales que toma en cuenta aspectos particulares del dominio. Los datos identificados como fuente para el modelo corresponden a datos hidrometeorológicos provenientes de redes de monitoreo ambiental.

Se profundizó en la revisión de tecnologías asociadas a *Big Data* y en su aplicación al dominio específico de los datos ambientales. Con esto se logró identificar las diferentes estrategias para la captura, almacenamiento y tratamiento de los datos desde este enfoque y a su vez se encontró que dependen totalmente de la naturaleza de los datos que se quieran estudiar.

Se definió y detalló un modelo por capas que permite hacer la administración y análisis de datos ambientales (hidrometeorológicos) mediante el planteamiento de tres capas, una de almacenamiento, otra de análisis y una final de consumo, en las cuales se pueden incluir tecnologías tradicionales y las asociadas a *Big Data*.

Se hizo una validación del modelo en un caso de estudio particular, tomando datos hidrometeorológicos suministrados por el Instituto de Estudios Ambientales – IDEA – de la Universidad Nacional de Colombia Sede Manizales y estructurando una arquitectura particular. Se realizaron análisis mediante técnicas asociadas a *Big Data* e Inteligencia Artificial, representados en predicciones de comportamientos futuros de las variables precipitación, temperatura y presión barométrica, las cuales quedan sujetas a validación y agrupamiento con algoritmos de Clustering usando *Mahout* y *Hadoop*.

6.2 Trabajos futuros

Como trabajos futuros se plantean diferentes aspectos enmarcados dentro de las diferentes capas del modelo.

Para para la capa de almacenamiento se plantea trabajar en la aplicación de una arquitectura que comprenda la integración de herramientas para datos estructurados y no estructurados al tiempo. Además, se espera, poder incluir nuevas fuentes de datos para

contar con mayor información del comportamiento de las variables y entregar otro tipo de análisis o predicciones.

Para la capa de análisis, se puede ampliar a otras técnicas que trabajen sobre otros atributos o que presenten otro tipo de resultados. También se plantean mejoras para los diferentes métodos aplicados.

- Para las predicciones se plantea revisar otros tipos de arquitecturas de Deep learning como las redes profundas basadas en máquinas restringidas de Boltzmann (Larochelle & Bengio, 2008) o utilizando múltiples autencoders apilados. También se considera incluir nuevos atributos como entradas del autoncoder y de la red.
- Para los análisis por medio de *Hadoop* y *Mahout*, los cuales fueron realizados en un nodo simple, se plantea la posibilidad de implementarlos para múltiples nodos, con el fin de aprovechar los beneficios de la distribución de tareas y comparar los tiempos de respuesta y eficiencia de los algoritmos. A su vez, se puede ampliar los algoritmos utilizados y revisar otras tecnologías de análisis acopladas a lenguajes de programación como R o Python.

Para la capa de consumo se plantea la creación de interfaces que puedan llevar de forma más amigable los resultados de los análisis y predicciones a los usuarios expertos en el campo ambiental y climático, para que ellos puedan emplearlos en procesos de toma de decisiones o generación de alertas. Teniendo en cuenta otras técnicas de análisis que se describen en el marco conceptual, se plantea también realizar procesos de análisis multidimensional, específicamente con herramientas OLAP.

Bibliografía

- Abhishek, K., Kumar, A., Ranjan, R., & Kumar, S. (2012). A rainfall prediction model using artificial neural network. En *2012 IEEE Control and System Graduate Research Colloquium (ICSGRC)* (pp. 82-87). <http://doi.org/10.1109/ICSGRC.2012.6287140>
- Adell, F., & Guersenzvaig, A. (2013). «*Big Data*» y los nuevos métodos de visualización de la información.
- Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep Machine Learning - A New Frontier in Artificial Intelligence Research. *IEEE Computational Intelligence Magazine*, 5(4), 13-18. <http://doi.org/10.1109/MCI.2010.938364>
- Barbierato, E., Gribaudo, M., & Iacono, M. (2014). Performance evaluation of NoSQL big-data applications using multi-formalism models. *Future Generation Computer Systems*, 37, 345-353. <http://doi.org/10.1016/j.future.2013.12.036>
- Bartok, J., Habala, O., Bednar, P., Gazak, M., & Hluchý, L. (2010). *Data Mining* and integration for predicting significant meteorological phenomena. *Procedia Computer Science*, 1(1), 37-46. <http://doi.org/10.1016/j.procs.2010.04.006>
- Beltrán-Castro, J., Valencia-Aguirre, J., Orozco-Alzate, M., Castellanos-Domínguez, G., & Travieso-González, C. M. (2013). Rainfall Forecasting Based on Ensemble Empirical Mode Decomposition and Neural Networks. En I. Rojas, G. Joya, & J. Gabestany (Eds.), *Advances in Computational Intelligence* (pp. 471-480). Springer Berlin Heidelberg. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-38679-4_47
- Bu, Y., Howe, B., Balazinska, M., & Ernst, M. D. (2010). HaLoop: efficient iterative data processing on large clusters. *Proceedings of the VLDB Endowment*, 3(1-2), 285–296.

Bustamante Martínez, A., Galvis Lista, E. A., & Gómez Flórez, L. C. (2013). Técnicas de modelado de procesos de ETL: una revisión de alternativas y su aplicación en un proyecto de desarrollo de una solución de BI. (Spanish). *ETL Processes modeling techniques: an alternatives review and its application in a BI solution development project. (English)*, 18(1), 185-191.

Chaudhuri, S., & Dayal, U. (1997). An Overview of Data Warehousing and OLAP Technology. *SIGMOD Rec.*, 26(1), 65–74. <http://doi.org/10.1145/248603.248616>

Chen, C. L. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on *Big Data*. *Information Sciences*, 275, 314-347. <http://doi.org/10.1016/j.ins.2014.01.015>

Chen, M., Mao, S., & Liu, Y. (2014). *Big Data: A Survey*. *Mobile Networks and Applications*, 19(2), 171-209. <http://doi.org/10.1007/s11036-013-0489-0>

Cloquell, V., Santamarina, M. C., & Hospitaler, A. (2001). Nuevo procedimiento para la normalización de valores numéricos en la toma de decisiones. Presentado en XVII Congreso Nacional de Ingeniería de Proyectos, Murcia. Recuperado a partir de <http://www.unizar.es/aeipro/finder/ORGANIZACION%20Y%20DIRECCION/DD18.htm>

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal Of Big Data*, 2(1), 1–24.

Dash, R., & Dash, P. K. (2016). A Hybrid Stock Trading Framework Integrating Technical Analysis with Machine Learning Techniques. *The Journal of Finance and Data Science*. Recuperado a partir de <http://www.sciencedirect.com/science/article/pii/S2405918815300179>

Dean, J., & Ghemawat, S. (2008). *MapReduce: Simplified Data Processing on Large Clusters*. *Commun. ACM*, 51(1), 107–113. <http://doi.org/10.1145/1327452.1327492>

Dhanya, C. T., & Nagesh Kumar, D. (2009). *Data Mining for Evolving Fuzzy Association Rules for Predicting Monsoon Rainfall of India*. *Journal of Intelligent Systems*, 18(3), 193-210.

Dittrich, J., Quiané-Ruiz, J.-A., Jindal, A., Kargin, Y., Setty, V., & Schad, J. (2010). *Hadoop++: making a yellow elephant run like a cheetah (without it even noticing)*. *Proceedings of the VLDB Endowment*, 3(1-2), 515–529.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.

Duque, N., Hernández, E., Pérez, Á., Arroyave, A., & Espinosa, D. (2016). Modelo para el proceso de extracción, transformación y carga en almacenes de datos. Una aplicación con datos ambientales. En edición. *Revista Ciencia e Ingeniería Neogranadina*, 26(2).

Duque Méndez, N. D., Orozco Alzate, M., & Hincapié, L. (2011). Minería de Datos para el Análisis de Datos Meteorológicos. En *Tendencias en Ingeniería de Software e Inteligencia Artificial*, 4, 105-114.

Duque Méndez, N. D., Vélez Upegui, J. J., & Orozco Alzate, M. (2015). Análisis multidimensional de datos ambientales. En *Entendimiento de fenómenos ambientales mediante análisis de datos* (Primera, pp. 115-132). Manizales, Colombia: Universidad Nacional de Colombia -Sede Manizales.

Duque-Méndez, N. D., Orozco-Alzate, M., & Vélez, J. J. (2014). Hydro-meteorological data analysis using OLAP techniques. *DYNA*, 81(185), 160-167.
<http://doi.org/10.15446/dyna.v81n185.37700>

ETESA. (2009). Duración media de brillo solar u horas de sol. Recuperado a partir de http://www.hidromet.com.pa/brillo_solar.php

Galbraith, J. R. (2014). Organization Design Challenges Resulting from *Big Data*. *Journal of Organization Design*, 3(1), 2-13. <http://doi.org/10.7146/jod.3.1.8856>

Gandomi, A., & Haider, M. (2015). Beyond the hype: *Big Data* concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144. <http://doi.org/10.1016/j.ijinfomgt.2014.10.007>

George, G., Haas, M. R., & Pentland, A. (2014, abril). *Big Data* and Management. *Academy of Management Journal*, pp. 321-326.

Grossman, R. L., Kamath, C., Kegelmeyer, P., Kumar, V., & Namburu, R. (2013). *Data Mining for Scientific and Engineering Applications*. Springer Science & Business Media.

Grover, A., Kapoor, A., & Horvitz, E. (2015). A Deep Hybrid Model for Weather Forecasting (pp. 379-386). Presentado en International Conference on Knowledge Discovery and *Data Mining* KDD'15, Sydney. <http://doi.org/http://dx.doi.org/10.1145/2783258.2783275>

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3.^a ed.). Elsevier.

Hernández, E., Sánchez-Anguix, V., Julian, V., Palanca, J., & Duque, N. (2016). Rainfall prediction: A Deep Learning approach (pp. 151-163). Presentado en Proceedings 11th International Conference on Hybrid Artificial Intelligence Systems.

Ingersoll, G. (2009, septiembre 8). Introducing Apache *Mahout*. Recuperado 9 de mayo de 2016, a partir de <http://www.ibm.com/developerworks/java/library/j-Mahout/>

Ingersoll, G. (2012, febrero 13). Apache *Mahout*: Aprendizaje escalable con máquina para todos. Recuperado 9 de mayo de 2016, a partir de <http://www.ibm.com/developerworks/ssa/library/j-Mahout-scaling/>

Jaramillo Valbuena, S., & Londoño, J. M. (2015). Sistemas para almacenar grandes volúmenes de datos. *Revista Gerencia Tecnológica Informática*, 13(37), 17-28.

Bibliografía

Javlin Inc. (2015). CloverETL Rapid Data Integration. Recuperado 22 de enero de 2016, a partir de <http://www.cloveretl.com/products/community-edition>

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and Challenges of *Big Data* Research. *Big Data Research*, 2(2), 59-64. <http://doi.org/10.1016/j.bdr.2015.01.006>

Kusiak, A., Wei, X., Verma, A. P., & Roz, E. (2013). Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4), 2337-2342. <http://doi.org/10.1109/TGRS.2012.2210429>

Lämmel, R. (2008). Google's *MapReduce* programming model—Revisited. *Science of computer programming*, 70(1), 1–30.

Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42, 11-24. <http://doi.org/10.1016/j.patrec.2014.01.008>

Larochelle, H., & Bengio, Y. (2008). Classification Using Discriminative Restricted Boltzmann Machines. En *Proceedings of the 25th International Conference on Machine Learning* (pp. 536–543). New York, NY, USA: ACM. <http://doi.org/10.1145/1390156.1390224>

Larose, D. T., & Larose. (2014). *Discovering knowledge in data: an introduction to Data Mining* (Segunda). John Wiley & Sons. Recuperado a partir de <https://books.google.com.co/books?hl=es&lr=&id=UGu8AwAAQBAJ&oi=fnd&pg=PT22&dq=Discovering+Knowledge+in+Data:+An+Introduction+to+Data+Mining&ots=zrsQjcRMtN&sig=zQmcyonuyHALjoH0VUfzQrtEoi4>

Li, K.-C., Jiang, H., Yang, L. T., & Cuzzocrea, A. (2015). *Big Data: Algorithms, Analytics, and Applications*. CRC Press. Recuperado a partir de <https://books.google.com.co/books?hl=es&lr=&id=yIG3BgAAQBAJ&oi=fnd&pg=PP1&dq=>

Big+Data+Algorithms,+analytics+and+applications&ots=PGpvGqMNS&sig=Uqxr115FFstt9djCNjIYGxHEVzw

Lin, W.-Y., Hu, Y.-H., & Tsai, C.-F. (2012). Machine learning in financial crisis prediction: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), 421–436.

LISA lab. (2015). Deep Learning Tutorials — DeepLearning 0.1 documentation. Recuperado 20 de octubre de 2015, a partir de <http://deeplearning.net/tutorial/index.html>

Liu, J. N. , Hu, Y., You, J. J., & Chan, P. W. (2014). Deep Neural Network Based Feature Representation for Weather Forecasting (pp. 261-267). Presentado en International Conference on Artificial Intelligence. Recuperado a partir de <http://worldcomp-proceedings.com/proc/p2014/ICA3405.pdf>

Liu, J. N. K., Li, B. N. L., & Dillon, T. S. (2001). An improved Naive Bayesian classifier technique coupled with a novel input solution method. *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews*, 31(2), 249-256. <http://doi.org/10.1109/5326.941848>

Luk, K. C., Ball, J. E., & Sharma, A. (2000). A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting. *Journal of Hydrology*, 227(1–4), 56-65. [http://doi.org/10.1016/S0022-1694\(99\)00165-1](http://doi.org/10.1016/S0022-1694(99)00165-1)

Maino, D. G., Uzal, L., & Granitto, P. M. (2014). Predicción de sistemas dinámicos con redes neuronales profundas (pp. 107-114). Presentado en XLIII Jornadas Argentinas de Informática e Investigación Operativa (43JAIIO)-XV, Buenos Aires. Recuperado a partir de <http://hdl.handle.net/10915/41734>

Michalski, R. S., Bratko, I., & Bratko, A. (Eds.). (1998). *Machine Learning and Data Mining; Methods and Applications*. New York, NY, USA: John Wiley & Sons, Inc.

Bibliografía

Microsoft. (2014, junio). Developing *Big Data* solutions on Microsoft Azure HDInsight. Recuperado 27 de mayo de 2016, a partir de <https://msdn.microsoft.com/en-us/library/dn749874.aspx>

Ocampo López, O. L., & Vélez Upegui, J. J. (2015). Análisis climatológico para el departamento de Caldas. En *Entendimiento de fenómenos ambientales mediante análisis de datos* (Primera, pp. 1-44). Manizales, Colombia: Universidad Nacional de Colombia - Sede Manizales.

O'Leary, D. E. (2013). Artificial Intelligence and *Big Data*. *IEEE Intelligent Systems*, 28(2), 96-99.

Organización Mundial de Meteorología OMM. (2008). Guía de Instrumentación Meteorológica y Métodos de Observación.

Orozco-Alzate, M., Velez-Upegui, J. J., & Duque-Mendez, N. D. (2014). Data Acquisition for Hydrometeorological Monitoring. *IEEE Potentials*, 33(5), 22-28. <http://doi.org/10.1109/MPOT.2013.2292534>

Pentaho. (2016). Data Integration | Pentaho Community. Recuperado 22 de enero de 2016, a partir de <http://community.pentaho.com/projects/data-integration/>

Portugal, I., Alencar, P., & Cowan, D. (2015). The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review. *arXiv*. Recuperado a partir de <http://arxiv.org/abs/1511.05263>

Prekopcsák, Z., Makrai, G., Henk, T., & Gaspar-Papanek, C. (2011). Radoop: Analyzing *Big Data* with rapidminer and *Hadoop*. En *Proceedings of the 2nd RapidMiner community meeting and conference (RCOMM 2011)* (pp. 865–874). Citeseer. Recuperado a partir de <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.296.5278&rep=rep1&type=pdf>

Ramírez Romero, T. A., Patiño Ortiz, J., & Patiño Ortiz, M. (2015). Comparativa entre Data Warehouse y generación dinámica de consultas en SQL. Presentado en 15° Congreso

Nacional de Ingeniería Electromecánica y de Sistemas (CNIES), México, D.F. Recuperado a partir de <http://www.sepi.esimez.ipn.mx/cnies/memorias/SIS01.pdf>

Rodríguez Jiménez, R. M., Capa, Á. B., & Portela Lozano, A. (2004). *Meteorología y Climatología*. Madrid: Fundación Española para la Ciencia y la Tecnología. Recuperado a partir de <http://cab.inta-csic.es/uploads/culturacientifica/adjuntos/20130121115236.pdf>

Sagiroglu, S., & Sinanc, D. (2013). *Big Data: A review*. En *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42-47). <http://doi.org/10.1109/CTS.2013.6567202>

Sahai, A. K., Soman, M. K., & Satyan, V. (2000). All India summer monsoon rainfall prediction using an artificial neural network. *Climate Dynamics*, 16(4), 291-302. <http://doi.org/10.1007/s003820050328>

Sawale, G. J., & Gupta, S. R. (2013). Use of Artificial Neural Network in *Data Mining For Weather Forecasting*. *International Journal Of Computer Science And Applications*, 6(2), 383-387.

Shi, D., Lee, Y., Duan, X., & Wu, Q. H. (2001). Power system data warehouses. *IEEE Computer Applications in Power*, 14(3), 49-55. <http://doi.org/10.1109/MCAP.2001.952937>

Silberschatz, A., Korth, H. F., Sudarshan, S., Pérez, F. S., Cordero, A. G., & Fernández, J. C. (2002). *Fundamentos de bases de datos*. McGraw-Hill. Recuperado a partir de <http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=SIDINA.xis&method=post&formato=2&cantidad=1&expresion=mfn=003016>

Simo, B., Habala, O., Tran, V., Krammer, P., & Hluchy, L. (2011). Using ADMIRE *Data Mining* and integration tools in hydrological forecast use case. En *2011 15th IEEE International Conference on Intelligent Engineering Systems (INES)* (pp. 215-220). <http://doi.org/10.1109/INES.2011.5954747>

Bibliografía

Sistema de información Ambiental de Colombia - SIAC. (2011). Recuperado 2 de septiembre de 2015, a partir de <https://www.siac.gov.co/portal/default.aspx>

Song, J., Guo, C., Wang, Z., Zhang, Y., Yu, G., & Pierson, J.-M. (2015). HaoLap: A *Hadoop* based OLAP system for *Big Data*. *Journal of Systems and Software*, 102, 167-181. <http://doi.org/10.1016/j.jss.2014.09.024>

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., ... Robinson, G. E. (2015). *Big Data*: astronomical or genomics? *PLoS Biol*, 13(7), e1002195.

Talburt, J. R., & Zhou, Y. (2015). *Entity Information Life Cycle for Big Data: Master Data Management and Information Integration*. Morgan Kaufmann. Recuperado a partir de <https://books.google.com.co/books?hl=es&lr=&id=Td-cBAAAQBAJ&oi=fnd&pg=PP1&dq=Entity+Information+Life+Cycle+for+Big+Data&ots=vnXHKNj4Ah&sig=1FTiwgZhFXrBAQ1ckeWTgkBSS4s>

Talend. (2016). Application Integration Solutions & ESB Platform from Talend. Recuperado 22 de enero de 2016, a partir de <http://www.talend.com/products/application-integration>

Tamayo, M., & Moreno, F. J. (2006). Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP. *Ingeniería e Investigación*, 26(3), 135-142.

The Apache Software Foundation. (2016a). Apache *Mahout*: Scalable machine learning and *Data Mining*. Recuperado 6 de mayo de 2016, a partir de <http://Mahout.apache.org/>

The Apache Software Foundation. (2016b). Welcome to Apache™ *Hadoop*®! Recuperado 4 de mayo de 2016, a partir de <http://Hadoop.apache.org/>

Toth, E., Brath, A., & Montanari, A. (2000). Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology*, 239(1–4), 132-147. [http://doi.org/10.1016/S0022-1694\(00\)00344-9](http://doi.org/10.1016/S0022-1694(00)00344-9)

Valverde Ramírez, M. C., de Campos Velho, H. F., & Ferreira, N. J. (2005). Artificial neural network technique for rainfall forecasting applied to the São Paulo region. *Journal of Hydrology*, 301(1–4), 146-162. <http://doi.org/10.1016/j.jhydrol.2004.06.028>

Villanueva Chávez, J. (2011). *Marco de trabajo basado en ontologías para el proceso ETL* (Tesis de maestría). Instituto Politécnico Nacional, México, D.F. Recuperado a partir de <http://webserver.cs.cinvestav.mx/TesisGraduados/2011/TesisJoelVillanueva.pdf>

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and Composing Robust Features with Denoising Autoencoders. En *Proceedings of the 25th International Conference on Machine Learning* (pp. 1096–1103). New York, NY, USA: ACM. <http://doi.org/10.1145/1390156.1390294>

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). *Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering*, 26(1), 97-107. <http://doi.org/10.1109/TKDE.2013.109>