



## **Ingeniería en Computación**

### **Bases de Datos II**

#### **Tarea Programada II: Caeli**

#### **Profesor:**

Erick Hernández Bonilla

#### **Estudiantes:**

Andrea Abarca Baltodano (2015088898)

Samantha Arburola León (2013101697)

Victor Chaves Díaz (2015107095)

Kevin Lobo Chinchilla (2015088135)

#### **II Semestre**

**30 de Octubre, 2016**

## **Tabla de Contenido**

<b>Tabla de Contenido</b>	<b>2</b>
<b>Introducción</b>	<b>3</b>
<b>Descripción del Proyecto</b>	<b>4</b>
<b>Arquitectura</b>	<b>5</b>
Web Crawler	5
Hadoop	6
Base de Datos	8
Aplicación Web	9
<b>Herramientas</b>	<b>10</b>
<b>Manual de Usuario</b>	<b>12</b>
<b>Conclusiones</b>	<b>21</b>
<b>Referencias</b>	<b>22</b>

# Introducción

En este proyecto veremos desarrollado un ejemplo de minería de datos así como su respectivo procesamiento en Hadoop aplicando un mapeo y reducción junto con cálculos de promedio por las diversas características climáticas encontradas en el sitio web en [tutiempo.net/climate](http://tutiempo.net/climate).

Para obtener los datos se elabora un algoritmo de extracción para cada una de las secciones encontradas dentro de los registros del clima para cada año recordando que en cada país hay numerosas estaciones meteorológicas. Esto con la implementación de un Web Crawler desarrollado a la medida, el cual visitará sólo las páginas web incumbentes evitando el sobre trabajo.

Este resultado se procesa con jobs de Hadoop los cuales son responsables de mapear la información y producir archivos con los resultados en un formato específico. Cada uno de esos archivos son introducidos a la base de datos mediante un programa en Java donde se relacionan los datos como países y estaciones a los promedios y años en los cuales fueron obtenidos.

# Descripción del Proyecto

El presente proyecto consta de un sitio web estadístico, que muestra los resultados del clima mundial:

- Los 10 países con los máximos promedios generales.
- Los 10 países con los mínimos promedios generales.
- Para cada país el año en que cada uno de las variables fue la máxima.
- Para cada país el año en que cada uno de las variables fue la mínima.
- El promedio de temperatura para cada continente, en grupos de 10 años.
- Por país la estación que tiene los valores máximos.
- Por país la estación que tiene los valores mínimos.
- Por continente los países con los valores máximos.
- Por continente los países con los valores mínimos.

Esta información se extrae de una base de datos que contiene la información generada en Hadoop, desde un almacenamiento escrito por el Web Crawler, este último se encarga de extraer datos del sitio <http://en.tutiempo.net/climate> que son información de continentes, países y estaciones; además, por cada estación, las siguientes estadísticas:

- Año
- Promedio de temperatura anual
- Promedio de temperatura máxima anual
- Promedio de temperatura mínima anual
- Precipitación de lluvia o nieve anual
- Promedio anual de velocidad del viento
- Días con lluvia al año
- Días con nieve
- Días con tormenta
- Días con niebla
- Días con tornado
- Días con granizo

# Arquitectura

## Web Crawler

**PHP:** Fue el lenguaje mediante el cual se programó el Web Crawler en su versión 7. Posee funciones para facilitar el acceso a los documentos de las páginas web, las cuales son utilizadas por la biblioteca Simple DOM HTML.

**Biblioteca Simple DOM HTML:** Biblioteca creada por S.C. Chen basada en la idea de Jose Solorzano y con contribuciones por parte de Yousuke Kumakura. Provee funciones para acceder directamente a los elementos de html y sus tags. De esta biblioteca se utilizaron las siguientes funciones:

- **file\_get\_html():** Recibe como parámetro el link de la página web de la cual se desea obtener el html. Retorna el html de la página si está disponible.
- **find():** Se le aplica al resultado de file\_get\_html(). Recibe como parámetro el tag que se desea buscar. Por ejemplo `html->find('a')` encuentra todos los tags a de un documento html y devuelve un array con sus ocurrencias.
- **plaintext:** Devuelve el texto plano de un tag, es decir, el innerHTML de este. Por ejemplo, `<p>Hola</p>->plaintext` retorna Hola.
- **href:** Retorna el atributo href del elemento html al cual se le aplique.
- **parent():** Retorna el elemento html padre del elemento html al cual se le aplica.
- **children():** Retorna el elemento html hijo del elemento html al cual se le aplica.

**Algoritmo implementado:** El Web Crawler recibe la dirección de la página web: <http://en.tutiempo.net/climate/> a partir de esta inicia la búsqueda de continentes mediante el uso de la función find() y del atributo title de los tags a que poseen los países. Luego por cada continente se accede a sus países con el mismo método utilizando la función find().

Posteriormente por cada país se accede a su estación de la misma forma en la que se accedió a los elementos anteriores y al acceder a una estación se verifica que posean datos correctos, de lo contrario se omite su inserción. Dado que en algunas ocasiones, los países tienen más de una página de estaciones, la página siguiente se busca, se accede a ella y se llama recursivamente a la función que chequea estaciones.

## Hadoop

Para Hadoop se crean 9 jobs diferentes, uno por cada cálculo especificado.

Para los puntos *c*, *d*, *f*, *g*, *h* e *i* se utiliza un tipo de dato propio que contiene los 11 valores y:

- El año (para los puntos *c* y *d*)
- La estación (para los puntos *f* y *g*)
- El país (Para los puntos *h* e *i*)

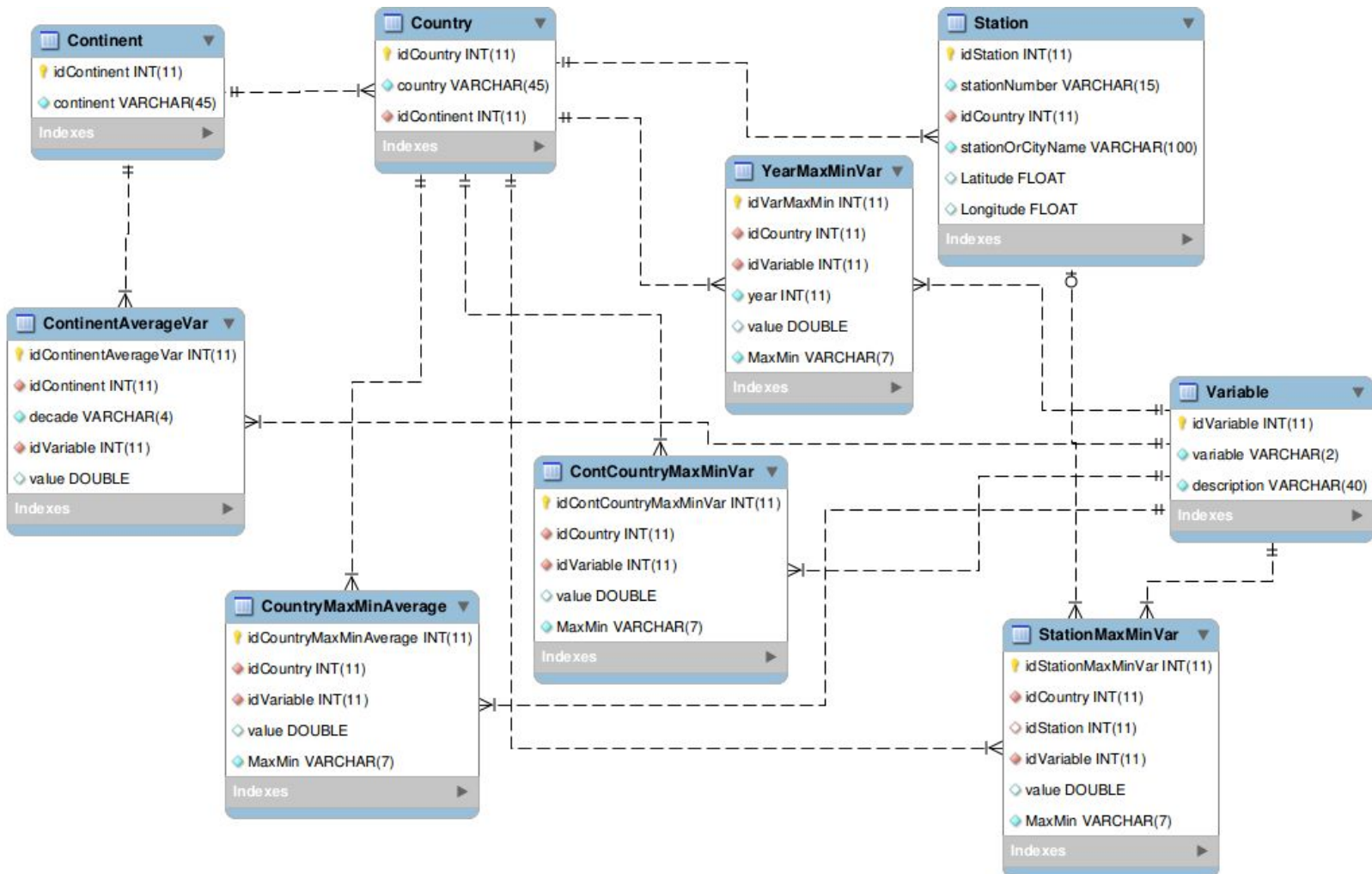
Donde el trabajo del mapper consiste en crear el objeto descrito, donde valores no existentes (representados por un guión) se convierten en el valor máximo y mínimo que puede guardar un `float`, para los que calculen el valor mínimo y máximo respectivamente. El reducer mantiene una serie de variables para almacenar el valor máximo/mínimo actual, y comienza a comparar la variable almacenada en el valor actual que se está procesando, esto por cada una de las 11 variables, dando al final los valores máximos/mínimos y el año/estación/país correspondiente para cada variable, posteriormente creando un nuevo objeto con estos valores y escribiendolos al output final.

Para los tops (puntos *a* y *b*) y el promedio por grupos de 10 años (punto *e*), el proceso involucrado es similar entre los tres: calcular los promedios consiste en tomar todos los valores de las variables asociadas a una llave, sumarlos y contar aquellos que forman parte de esta suma, para poder realizar la división (total sumado / total contados).

Específicamente para los tops se hace uso de TreeMaps, estructuras de datos que acomodan los datos en orden correspondiendo a la llave, en este caso se utiliza el valor asociado a una variable promediada como llave y el país como el valor. Una vez que existen más de 10 elementos en el TreeMap, se elimina una pareja. Para poder recolectar estos valores se hace uso de la función de `Reducer.cleanup`, que se ejecuta al final de las llamadas de `Reducer.reduce`.

# Base de Datos

Para la creación de la base se utilizó MySQL como motor y el modelo relacional trabajado con MySQL Workbench se diseñó de la siguiente manera:





## Aplicación Web

La aplicación web se utiliza para acceder a los datos obtenidos mediante Hadoop. Para su creación se utilizó:

**PHP:** Como código que se ejecuta en el servidor. Se utiliza el driver mysqli para conectar con la base de datos y obtener la información requerida.

**JavaScript y jQuery:** Se utiliza JavaScript como código que se ejecuta en el cliente junto a su biblioteca jQuery. Se utilizó ajax para obtener la información mediante php.

**API DataTables.net:** Brinda funciones y facilidades para el manejo de tablas HTML.

Esta aplicación web se divide en 5 secciones:

- **Top 10 Países - Promedios:** Muestra filtros y la tabla para obtener y mostrar los 10 países con los máximos promedios generales de las variables y con los mínimos promedios generales.
- **País - Año - Min - Max:** Muestra filtros y la tabla para obtener y mostrar para cada país el año en cada una de las variables fue la máxima y en el que cada una fue la mínima.
- **País - Estación - Min - Max:** Muestra filtros y la tabla para obtener y mostrar para cada país la estación que tiene los valores máximos de cada variables y la que tenga los valores mínimos.
- **Continente - Promedios - Décadas:** Muestra filtros y la tabla para obtener y mostrar el promedio de cada una de las variables por década, es decir, en grupos de 10 años.
- **Continente - País - Min - Max:** Muestra filtros y la tabla para obtener y mostrar por continente los países con los valores máximos de cada variable y los valores mínimos.

# Herramientas

## Eclipse

Para codear los jobs de Hadoop se usa Eclipse Neon, esto debido a que es un IDE que no consume tantos recursos y es más liviano como comúnmente se usa decir en el ambiente programacional.

Para su instalación seguir los pasos:

1. Descargar Eclipse Neon desde <https://www.eclipse.org/downloads/>
2. Extraer Eclipse en /opt/
3. Abrir la terminal desde la carpeta donde se descargó el archivo comprimido de Eclipse Neon

```
cd /opt/ && sudo tar -zxvf  
~/Downloads/eclipse-inst-linux64.tar.gz
```

Ejecute el archivo `eclipse-inst` y seleccione Eclipse IDE for Java Developers

4. Crean launcher con el comando:

```
sudo gksudo gedit /usr/share/applications/eclipse.desktop
```

5. En el archivo creado para el Launcher deben pegar

```
[Desktop Entry]  
Name=Eclipse Neon  
Type=Application  
(acá deben revisar la carpeta donde descomprimieron eclipse)  
Exec=/opt/eclipse/eclipse  
Terminal=false  
Icon=/opt/eclipse/icon.xpm  
Comment=Integrated Development Environment  
NoDisplay=false  
Categories=Development;IDE;  
Name[en]=Eclipse
```

**Biblioteca mysql-connector**

Esta biblioteca fue utilizada en su versión 5.1.40 para conectarse a la base de datos de MySQL desde java con el objetivo de guardar los datos obtenidos del map-reduce de Hadoop, por lo que se implementó la clase *Txt2sql* que recibe el archivo y el tipo de resultado correspondiente a las solicitudes especificadas en el documento de requerimiento del proyecto.

# Manual de Usuario

## Generalidades:

- **Filtros:** Permiten filtrar la información que se solicita, los filtros disponibles en la aplicación web son:
  - **Continente:** Permite seleccionar un continente para obtener la información únicamente de este.
  - **País:** Permite seleccionar un país para obtener la información únicamente de este. En caso de que haya un continente seleccionado, solo estarán disponibles los países de dicho continente para su selección.
  - **Variable:** Permite seleccionar una variable climática para obtener la información únicamente de esta. Las variables disponibles son:
    - T - Average annual temperature
    - TM - Annual average maximum temperature
    - Tm - Average annual minimum temperature
    - PP - Rain or snow precipitation total annual
    - V - Annual average wind speed
    - RA - Number of days with rain
    - SN - Number of days with snow
    - TS - Number of days with storm
    - FG - Number of foggy days
    - TN - Number of days with tornado
    - GR - Number of days with hail
  - **Max/Min:** Permite seleccionar si se desea mostrar la información que corresponde a los máximos, a los mínimos o ambos.
  - **Estación:** Permite seleccionar si se desea mostrar la información que corresponde a una estación meteorológica. Este filtro está disponible únicamente en la sección "País - Estación - Min - Max". Si se selecciona

un continente, permite seleccionar únicamente las estaciones de dicho continente. Si se selecciona un país, permite seleccionar únicamente las estaciones de ese país.

- **Década:** Permite seleccionar la década de la cual se desea mostrar la información. Filtro disponible únicamente en la sección "Continente - Promedios - Décadas".

### Manual:

Busque el siguiente icono en su escritorio y presione doble click sobre él. Se abrirá la web para observar los datos:



GENERAL

Top 10 Países - Promedios

País - Año - Min - Máx

País - Estación - Min - Máx

Continente - Promedios - Décadas

Continente - País - Min - Máx

Opciones de Ejecución

Los 10 países con los máximos/mínimos promedios generales.

Filtros

Continente

Todos

País

Todos

Variable

Todas

Min-Max

Ambos

Aplicar

Tabla

Copy

CSV

Excel

PDF

Print

Search:

Continente	País	Variable	Promedio	Max/Min
No data available in table				

Showing 0 to 0 of 0 entries

PreviousNext

Gentelella - Bootstrap Admin Template by Colorlib

Como puede observar, se le presenta una aplicación web mediante la cual puede acceder a los datos generados por Hadoop.

A un lado encontrará el SideBar con accesos a las diferentes páginas de la aplicación web, lo que le brinda acceso a otro tipo de información



La primera página permite filtrar y obtener la información de los diez países con los máximos y mínimos promedios generales de cada variable. Seleccione los filtros que desee y haga click en el botón "Aplicar". Se le mostrará la tabla con la información solicitada:

Los 10 países con los máximos/minimos promedios generales.

Filtros

Continente

Todos

País

Todos

Variable

Todas

Min-Max

Ambos

Aplicar

Tabla

CopyCSVPrint

Search:

Continente	País	Variable	Promedio	Max/Min
South America	Guyana	FG - Number of foggy days	109.72222	Max
North America	Saint Pierre and Miquelon	FG - Number of foggy days	108.18519	Max
Africa	Liberia	FG - Number of foggy days	99.57143	Max
Europe	Slovakia	FG - Number of foggy days	75.23185	Max
Europe	Germany	FG - Number of foggy days	71.14754	Max
Europe	Luxembourg	FG - Number of foggy days	69.6383	Max
Europe	Slovenia	FG - Number of foggy days	69.009094	Max
North America	United States	FG - Number of foggy days	68.85888	Max

Así mismo, desde la tabla puede utilizar el cuadro de texto de "Search" para filtrar los datos de la misma.

Los 10 países con los máximos/mínimos promedios generales.

Filtros

Continente

Todos

País

Todos

Variable

Todas

Min-Max

Ambos

Aplicar

Tabla

CopyCSVPrint

Search: Slovakia

Continente	País	Variable	Promedio	Max/Min
Europe	Slovakia	FG - Number of foggy days	75.23185	Max

Showing 1 to 1 of 1 entries (filtered from 220 total entries)

Previous1Next

Gentelella - Bootstrap Admin Template by Colorlib



Puede cambiar los filtros y presionar de nuevo el botón "Aplicar" y los datos de la tabla se actualizarán por los deseados.

Los 10 países con los máximos/minimos promedios generales.

Filtros

Continente

Africa

País

Angola

Variable

Todas

Min-Max

Ambos

Aplicar

Tabla

Copy

CSV

Print

Search:

Continente	País	Variable	Promedio	Max/Min
Africa	Angola	SN - Number of days with snow	0.027027028	Min

Showing 1 to 1 of 1 entries

Previous


1

Next

Gentelella - Bootstrap Admin Template by Colorlib

Las secciones "País - Año - Min - Max" y "Continente - País - Min - Max" se utilizan de manera análoga.

La sección "País - Estación - Min - Max" añade un nuevo filtro como se muestra a continuación. Este permite filtrar la información por estaciones meteorológicas:



Para cada país la estación en la que cada una de las variables fue la máxima/mínima

Filtros

Continente

Todos

País

Todos

Estación

Todos

Variable

Todas

Min-Max

Ambos

Aplicar

Tabla

Copy

CSV

Excel

PDF

Print

Continente	País	Nombre de Estación o Región	Número de Es
No data available in table			

Showing 0 to 0 of 0 entries

Previous

Next

Todos

589690 - 589690

A Coruña / Alvedro - 80020 (LECO)

A028 S. LAW DOME - 898120

AACHEN - 105010

Aalborg - 60300 (EKYT)

AARHUS SYD - 60740

AASIAAT MITTARFIA - 42021 (BGAA)

Abadan - 408310 (OIAA)

Abadeh - 408180 (OISA)

ABAG QI - 531920

Abakan - 287854 (UNAA)

ABAKAN - 298650 (UNKA)

ABASHIRI - 474090

Abbeville - 70050 (LFOI)

Abbotsford - 711080 (CYXX)

ABBOTSFORD AIRPORT - 741080

ABDALY - 405500

ABED - 61410

ABEE AGDM - 712850 (CXAF)

La sección "Continentes - Promedios - Décadas" elimina el filtro de países dado que no es necesario y brinda un nuevo filtro de décadas, que permite filtrar por décadas la información:

Por continente, el promedio de cada variable en grupos de diez años

Filtros

Continente

Todos

Década

Todas

Todas193019401950196019701980199020002010

Variable

Todas

Tabla

CopyCSVExcelPDFPrint

Search:

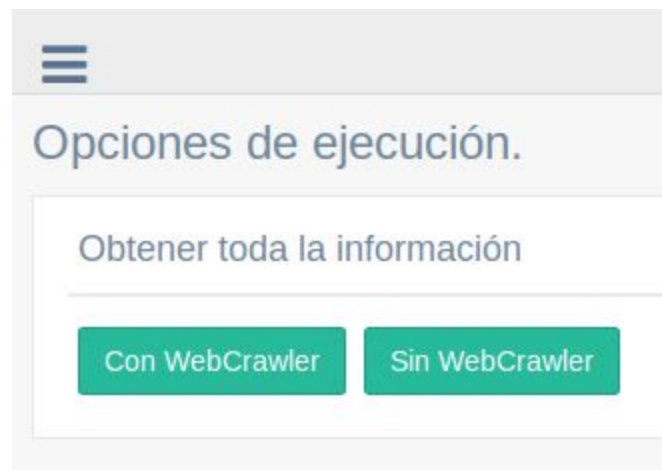
Continente	Década	Variable	Valor
No data available in table			

Showing 0 to 0 of 0 entries

PreviousNext

Gentelella - Bootstrap Admin Template by Colorlib

Finalmente, se brinda la sección "Opciones de Ejecución" en la cual se presentan dos botones:



- **Con WebCrawler:** Inicia primeramente la ejecución del WebCrawler antes de ejecutar los jobs de Hadoop. Luego de finalizar el WebCrawler, inicia la ejecución de jobs e inserción de resultados en la base de datos automáticamente.
- **Sin WebCrawler:** Igual al anterior pero sin la ejecución del WebCrawler. Se asume que existe el archivo data.txt generado por el WebCrawler.

# Conclusiones

La manera en que se extrajo la información para este proyecto evidencia lo expuesta que está esa información y como cualquiera con el tiempo para aprender a usar herramientas puede extraer datos sin ser detectados por su extracción sino viéndose como un visitante más. Esto lo podemos realizar de manera segura, es decir, claros de que no habrán represalias gracias al derecho constitucional que tienen todos los ciudadanos sobre la información pública.

Para programadores, todo es cuestión de un buen algoritmo y tiempo, esto último es el recurso indispensable para ejecutar el Web Crawler, este requiere una conexión estable y al ser desarrollado en PHP se evidencia como la inestabilidad de la red puede dañar procedimientos; algo que no se nota en las páginas web actuales ya que por medio del buscador o de protectores de interrupciones evitan que los usuarios finales no perciban las inconsistencias, ya que mientras esté trabajando en un hilo constante no habrá problema y de existir la interrupción se debe reiniciar el proceso, por esto es necesario de una función de *intentos* para resguardar la consistencia. El proceso para extraer y escribir los datos puede durar alrededor de 6 horas.

Por el lado del manejo de la información una vez adquirida es importante recalcar que celoso que es hadoop en sus procedimientos, ya que todo formato debe ir 100% acertivo para funcionar correctamente, de no ser así el sistema de archivos no trabajará óptimamente. El la interacción del programador con Hadoop es contrario al Web Crawler, ya que el primero tiene una curva de aprendizaje más grande que el segundo, el desarrollo de la solución es más extenso en el primero más en el segundo ese tiempo es el que se invierte en la extracción de los datos, ya que en tres días de trabajo se obtienen los jobs de hadoop más en 20 minutos estos dan los resultados.

En un proyecto de tan altas expectativas lograr el cien por ciento de funcionalidad y su automatización expande la visión de desarrollo para proyectar un profesional cada vez más completo.

# Referencias

M , Jim. (2016). *How to Install The Latest Eclipse in Ubuntu 16.04, 15.10 | UbuntuHandbook. Ubuntuhandbook.org*. Recuperado 15 de octubre de 2016, a partir de <http://ubuntuhandbook.org/index.php/2016/01/how-to-install-the-latest-eclipse-in-ubuntu-16-04-15-10/>

*PHP Simple HTML DOM Parser*. (2016). *Simplehtmldom.sourceforge.net*. Recuperado 26 October 2016, a partir de <http://simplehtmldom.sourceforge.net/>

*Apache Hadoop 2.7.2 – Hadoop: Setting up a Single Node Cluster..* (2016). *Hadoop.apache.org*. Recuperado 20 October 2016, a partir de <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html>

*MySQL :: MySQL Documentation*. (2016). *Dev.mysql.com*. Recuperado 9 October 2016, a partir de <https://dev.mysql.com/doc/>