

Arquitectura de un Almacén de Datos Espacial

Architecture of a Spatial Data Warehouse

Guillermo Pizarro ^{1,*}, Vanessa Jurado ¹, Shirley Coque ¹

INFORMACIÓN DEL ARTÍCULO

Fecha de recepción: 22 de marzo de 2018.

Fecha de aceptación: 4 de septiembre de 2018.

Resumen

Dada la necesidad de la gestión de datos operativos históricos geográficos, el presente trabajo propone y verifica el diseño y la implementación de una arquitectura para un Almacén de Datos Espacial (ADE), mediante el uso de herramientas FOSS4G como: PostgreSQL + PostGIS para la base de datos espacial; Talend Open Studio + SDI para el procedimiento ETL; GeoServer como servidor de mapas; SOLAP en SpagoBI Server y sus herramientas secundarias. La implementación del ADE (Almacén de Datos Espacial) permitió el análisis de un caso de estudio que demuestra la posibilidad de la aplicación de ADE en procesos de gestión de áreas sociales, de negocios o de investigación.

Palabras Clave:

Almacén de datos espacial, arquitectura, SpagoBI, PostgreSQL, PostGIS, GeoServer.

Clasificación JEL: O32.

Abstract

Given the need for historical geographic data management, this research proposes and verifies a design and implementation of an architecture for a Spatial Data Warehouse (SDW), by using FOSS4G tools such as: PostgreSQL + PostGIS for the base of spatial data; Talend Open Studio + SDI for the ETL procedure; GeoServer as a map server; SOLAP in SpagoBI Server and its secondary tools. The implementation of SDW allowed the analysis of a case study that demonstrates the possibility of the application of SDW in a management processes of social, business or research areas.

Keywords:

Spatial data warehouse, architecture, SpagoBI, PostgreSQL, PostGIS, GeoServer.

JEL Classification: O32.

¹ Universidad Politécnica Salesiana, Carrera de Ingeniería en Ciencias de la Computación, Grupo de Investigación en Ingeniería de Software e Ingeniería del Conocimiento, Guayaquil, Ecuador

* *Autor de correspondencia:*

Guillermo Pizarro, Universidad Politécnica Salesiana. Carrera de Ingeniería en Ciencias de la Computación. Chambers 227 y 5 de Junio. Guayaquil- Ecuador.
E-mail: gpizarro@ups.edu.ec. Tlf: (593-4) 259 0630

ENLACE DOI:

<http://dx.doi.org/10.31095/investigatio.2019.12.1>

Introducción

Los almacenes de datos proveen información histórica detallada de las operaciones de una empresa, para brindar apoyo en la toma de decisiones; sin embargo, al disponer de datos geográficos relacionado a esas operaciones, es necesario brindar otro tipo de servicio para visualizar esos datos en un mapa y tomar decisiones mejor informadas.

Con base en los trabajos publicados por Inmon y Kelley (1994) y Bohorquez (2003) propone una definición del Almacén de Datos Espacial (ADE), la cual cita lo siguiente: “*una colección de datos orientado a temas, integrada, variable en el tiempo, no volátil, que añade la geografía del dato en la base del análisis, para el apoyo de la toma de decisiones*”.

Existen trabajos relacionados a esta temática, haciendo uso del ADE en distintos campos de aplicación; tales como: i) visualización de áreas geográficas de intervención para la ejecución de proyectos de vinculación (Pizarro *et al.*, 2018); ii) detección de crímenes (Sale *et al.*, 2018); puntos calientes de biodiversidad (Hasanah y Trisminingsih, 2016); estudios hidro-ecológicos (Berrahou *et al.*, 2015); análisis de datos demográficos; (González y González, 2013); entre otros. Estos campos de aplicación hacen pertinente la propuesta del diseño de la arquitectura de un ADE, de acuerdo a las necesidades del usuario final.

Este trabajo de investigación se origina de un estudio descriptivo, detallado en Pizarro y Urvina (2017, pp. 33-35). Estos

autores realizan una comparación de ventajas y desventajas entre herramientas FOSS4G (*Free and Open Source Software for Geospatial*) que puedan ser utilizadas como un ADE, como *herramientas SOLAP: SpagoBI + GIS* (Uswatun y Trisminingsih, 2016) y *GeoMondrian* (Bogantes y Pandolfi, 2014); como *herramientas ETL automáticas: GeoKettle* en Uswatun y Trisminingsih, (2016) y *SDI de Talend Open Studio* en Oubennaceur (2011) y Fiedler (2011); hasta *ETL manual* mediante scripts (González y González, 2014).

El objetivo principal de esta investigación es la de proponer el diseño de una arquitectura para un ADE utilizando herramientas de código abierto y como objetivo secundario verificar su funcionamiento mediante el caso de estudio de la densidad geográfica de los cantones de la provincia del Guayas.

Materiales y Métodos

Este trabajo de investigación tiene un alcance descriptivo y hace uso del método experimental. Descriptivo, en cuanto a que se diseña un ADE, se lo detalla y se especifican los elementos que la componen. Experimental, debido a que se presenta una implementación detallada que puede ser replicada y se verifica la propuesta tecnológica mediante un caso de estudio.

Para llevar a cabo la propuesta del diseño de la arquitectura de un ADE, fue necesario:

1. Elaborar el diseño del Almacén de

Datos Espacial (ADE), mencionando los datos fuentes desde donde se alimentará el ADE, descripción de los componentes de la arquitectura y el detalle de la infraestructura.

2. Implementar el Almacén de Datos Espacial (ADE), presentando las herramientas a utilizar, la automatización del procedimiento ETL espacial, la puesta en marcha de un servidor de mapas y la visualización de los datos espaciales.

Diseño del Almacén de Datos Espacial (ADE)

Datos

Para elaborar el diseño del ADE es necesario considerar la utilidad o el para qué se utilizará un ADE. En este caso, se le utilizará como herramienta para visualizar la densidad geográfica de los cantones de la ciudad de Guayaquil.

Los datos fuentes son los del Censo de Población y Vivienda del 2010, 2001 y 1990; además de la cartografía de la provincia del

Guayas del 2010 y 2001, se obtuvo cartografía de la ciudad de Guayaquil de 1990. Los datos de estas fuentes fueron proporcionados por el INEC (Instituto Nacional de Estadística y Censo).

Componentes de la Arquitectura

En la Figura 1 se puede visualizar la arquitectura a utilizar para la implementación del ADE, además de la debida visualización de datos georreferenciados.

En la Arquitectura del ADE, se requieren los siguientes componentes:

i) Datos Heterogéneos

Los datos heterogéneos son aquellos que se encuentran en diferentes formatos (v.g. shapefile, SPSS, entre otros) o en diferentes motores de bases de datos (v.g. PostgreSQL, SQL Server, Oracle, entre otros). En este caso se tiene acceso a los siguientes elementos: a) Base de Datos (en *PostgreSQL*) de los proyectos de vinculación de la DTVCS; b) los archivos en formato *SPSS* de los Censos de

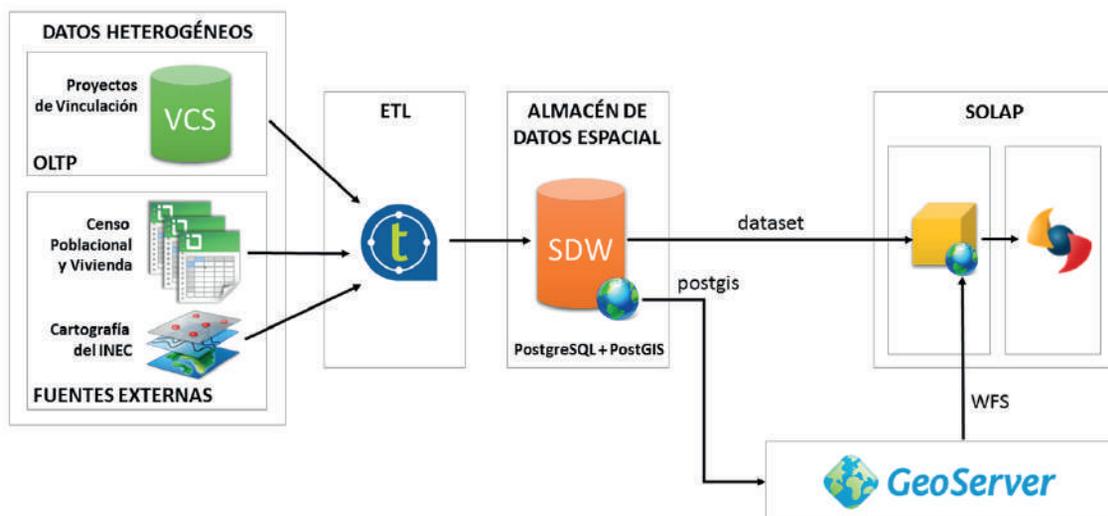


Figura 1. Arquitectura de un Almacén de Datos Espacial (ADE).

Población y Vivienda de los años 1990, 2001 y 2010; y c) los archivos en formato *shapefile* de la cartografía censal: de 1990 (*zonas y sectores de Guayaquil*), del 2001 y 2010 (*zonas, sectores, parroquias, cantones y provincia del Guayas*).

ii) Proceso ETL (Extraction – Transformation – Load)

En este procedimiento de extracción, transformación y carga se lo realizó con la herramienta de código abierto *Talend Open Studio (TOS)* incluyendo la extensión del *Spatial Data Integrator (SDI)* para poder trabajar con datos espaciales (Prunaire *et al.*, 2007). Con la herramienta de *TOS + SDI* se pudo extraer desde la cartografía la información espacial requerida para cargarla al ADE (*PostgreSQL + PostGIS*) mediante un minucioso procedimiento de transformación de reglas.

iii) Almacén de Datos Espacial (ADE)

En el ADE se almacena tanto la información escalar como la geográfica.

Se encuentra implementado en *PostgreSQL* haciendo uso de la extensión *PostGIS*.

iv) SOLAP (Spatial OnLine Analytical Processing)

La herramienta de procesamiento analítico en línea espacial facilita la visualización de los datos georreferenciados. Se utiliza el componente de *Location Intelligence de SpagoBI Server*; además, se alimenta de datos escalares a través de un dataset.

v) GeoServer

Extrae los datos espaciales de la cartografía almacenada en el ADE y ofrece el servicio de mapas a *SpagoBI Server* comunicándose vía *WFS (Web Feature Service)*, para visualizar la información geográfica deseada.

Infraestructura

La infraestructura del diseño de la arquitectura anterior, se visualiza en la siguiente Figura 2.

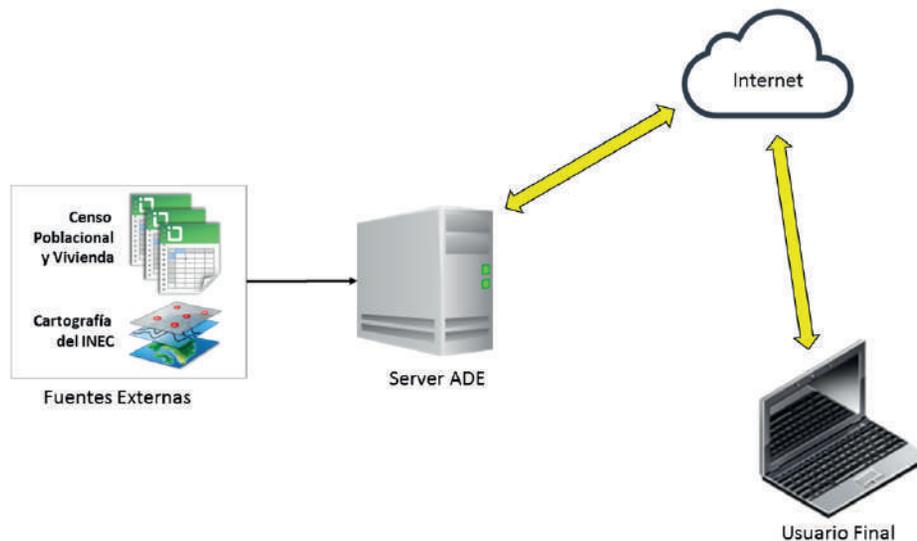


Figura 2. Infraestructura del ADE.

A continuación, el detalle de cada uno de los elementos de la infraestructura:

i) Fuentes externas

Dentro del diseño de la distribución de los servidores, este elemento se puede ubicar en un directorio dentro del Server ADE; sin embargo, se hace explícito en la Figura 2, debido a la importancia de estos datos que podrían actualizarse al existir un nuevo Censo de Población y Vivienda o de la cartografía censal.

ii) Server ADE

En este servidor se realizan las tareas del procedimiento ETL espacial y escalar para poblar al ADE de manera planificada y automática. También se encuentran: GeoServer y SpagoBI Server. Este servidor tiene conexión a internet para que los usuarios finales puedan acceder a la información deseada.

iii) Usuario Final

La persona que desee tener acceso al ADE, lo puede hacer mediante un navegador web y con una conexión a

Tabla 1.

Especificaciones técnicas del Servidor ADE.

internet.

Implementación del Almacén de Datos Espacial (ADE)

Especificaciones Técnicas del Server ADE

En la Tabla 1, se encuentran detalladas las especificaciones técnicas del Servidor ADE.

Para llevar a cabo el procedimiento ETL espacial fue necesario conocer acerca de la información almacenada en la cartografía proporcionada por el INEC, para esto se utilizó la herramienta QGIS, para la respectiva lectura de la información contenida en la cartografía. Una vez conocida la información en la cartografía se inició con el procedimiento ETL como se puede visualizar en la Figura 3, con la cartografía de las zonas, lo mismo se realiza con los sectores, parroquias, cantones y provincia.

El procedimiento ETL se realizó en

	Características	Descripción
Software	Sistema Operativo	Ubuntu 16.04 LTS (GPL y otras licencias libres)
	Base de Datos	PostgreSQL x64 9.5 + PostGIS (licencia PostgreSQL)
	Software para Inteligencia de Negocios	SpagoBI 5.1.0 (licencia pública Mozilla, GPL v3.0)
	Software ETL	Talend Open Studio (licencia Apache 2.0)
Hardware	Procesador	Intel Core i7
	Disco Duro	1TB.
	Memoria RAM	12GB.
	Conexión a Internet	mínimo 10MBPS dedicados.

Procedimiento ETL Espacial.

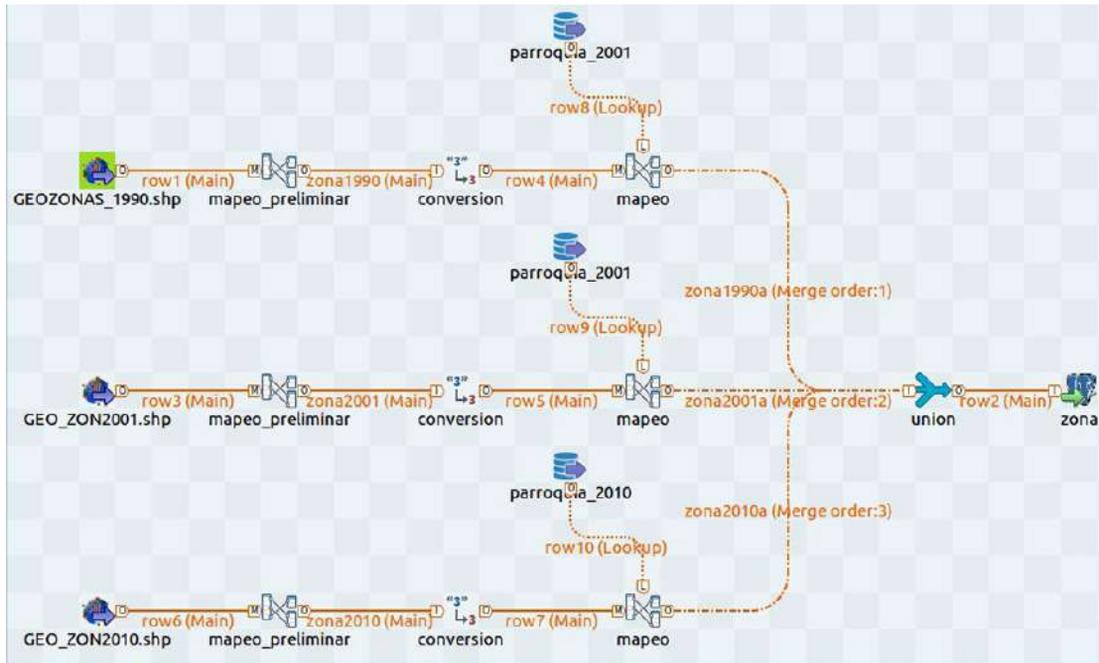


Figura 3. Procedimiento ETL de la cartografía censal de las zonas.

los siguientes pasos: 1) se accede al archivo; 2) se realiza un mapeo preliminar, como se puede verificar en la Figura 4; 3) se convierten algunas cadenas de caracteres en números; 4) se

hace el mapeo entre las zonas con las parroquias a las que pertenecen; 5) se unen los datos de 1990, 2001 y 2010; 6) se almacenan los datos espaciales en el ADE (Figura 4). Este mismo

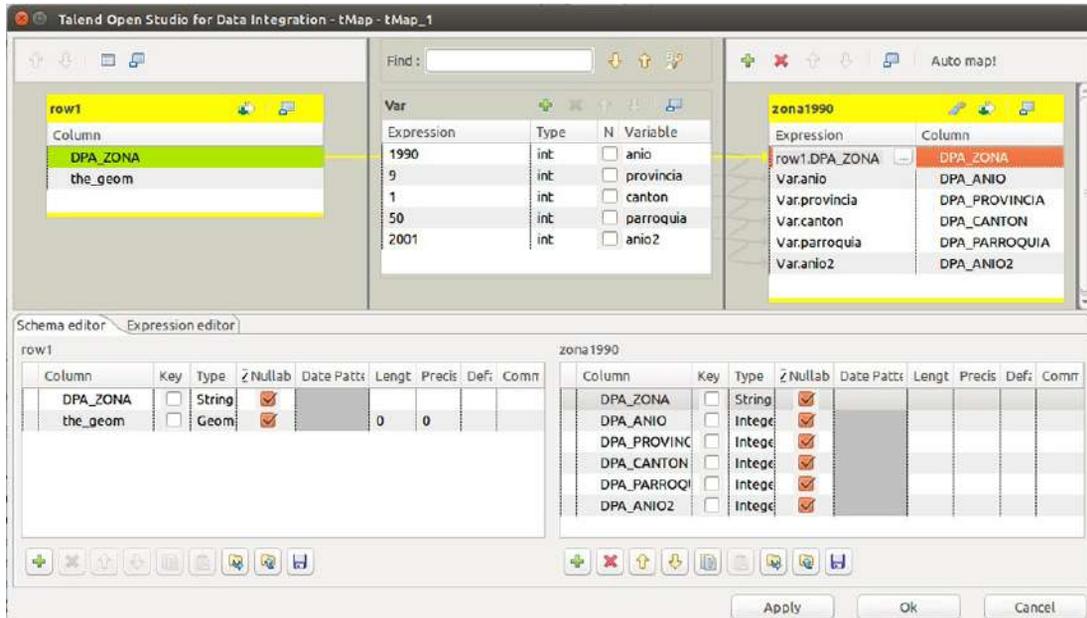


Figura 4. Mapeo preliminar de la cartografía.

procedimiento se aplicó a los datos escalares del Censo de Población y Vivienda de los años 1990, 2001 y 2010.

Servicio de mapas mediante Geoserver

Una vez cargado los datos en el ADE, se procede a la configuración del *GeoServer*, para que este servicio de mapas se pueda alimentar directamente de la cartografía cargada en el ADE. Para este caso se procede a extraer los mapas de los cantones de la provincia del Guayas del año 2010, lo mismo se hace si se desea extraer los mapas de las parroquias de la provincia del Guayas del año 2001, este tipo de configuraciones se debe realizar en el *GeoServer*.

Visualización de datos espaciales

Diseñada, implementada y en funcionamiento la arquitectura del ADE y configurado el *GeoServer*, se puede proceder a presentar datos espaciales mediante *SpagoBI Server* en su componente *Location Intelligence*.

Para su correcto funcionamiento es necesario realizar lo siguiente:

i) Crear un documento “*Generic Document*”, luego seleccionar el tipo “*Location Intelligence*”, después te va a mostrar dos opciones: “*Geo Engine*” y “*Gis Engine*”, la diferencia entre los dos es que el primero carga datos espaciales estáticos (hay que elaborar un mapa en SVG); en cambio, el segundo carga datos espaciales dinámicos desde *GeoServer*.

ii) Una vez seleccionado “*Gis Engine*”

se procede a asignarle un *datasource* que permite la conexión al ADE y un *dataset* al documento creado, que permite extraer datos desde el ADE.

iii) El *datasource* y el *dataset* deben ser creados antes de crear este documento. Es importante mencionar que el *dataset* debe contener una referencia que también exista en el WFS del mapa publicado en *GeoServer*, esto debido a que se deben enlazar los datos en *SpagoBI Server* y visualizar los mapas deseados.

iv) Una vez realizado lo anterior, se debe crear un archivo de configuración, para que se pueda visualizar de manera adecuada el mapa, a continuación el código del archivo *template.json*.

```
{
  "mapName": "Map",
  "analysisType": "choropleth",
  "indicatorContainer": "store",
  "storeType": "physicalStore",
  "indicators": [ ["total", "Total de Población"] ],
  "businessId": "ref_canton",
  "geoId": "fid",
  "analysisConf": {
    "method": "CLASSIFY_BY_QUANTILS",
    "fromColor": "#C0C0C0",
    "toColor": "#333399",
    "indicator": "total",
    "minRadiusSize": 2,
    "maxRadiusSize": 20
  },
  "targetLayerConf": {
    "text": "Canton
2010",
```

```

        "name": "2010_canton",
        "url": "http://localhost:8081/geoserver/wfs"
    },
    "selectMode": "cross",
    // 'cross' or 'detail'
    "crossnav": {
        "label": "CANTON_INEC_2010",
        "dynamicParams": [{"fid": "fid", "scope": "feature"}]
    },
    "role": "spagobi/admin",
    "lon": -79.99,
    "lat": -2.38,
    "zoomLevel": 4
}

```

Además, para vincular un reporte con los sectores / zonas / parroquias / cantones / provincia que se mostrarán en el mapa, se debe hacer uso de *SpagoBI Studio + iReport* para su debido diseño, elaboración e implementación. En el

archivo *template.json* se debe hacer referencia a este reporte para que se vincule con el mapa, como se puede verificar en la sección *crossnav*.

Resultados y discusión

En la Figura 6, se visualiza un mapa tematizado de acuerdo a la configuración descrita hasta el momento, donde se muestran los cantones de la provincia del Guayas del año 2010. Este mapa se muestra con colores degradados donde el color azul fuerte visualiza la mayor densidad poblacional; en cambio, el color gris implica una densidad poblacional menor.

Al seleccionar un Cantón, por ejemplo el Cantón Guayaquil (donde el color azul es más fuerte), se puede acceder a un reporte detallado de sus parroquias de los

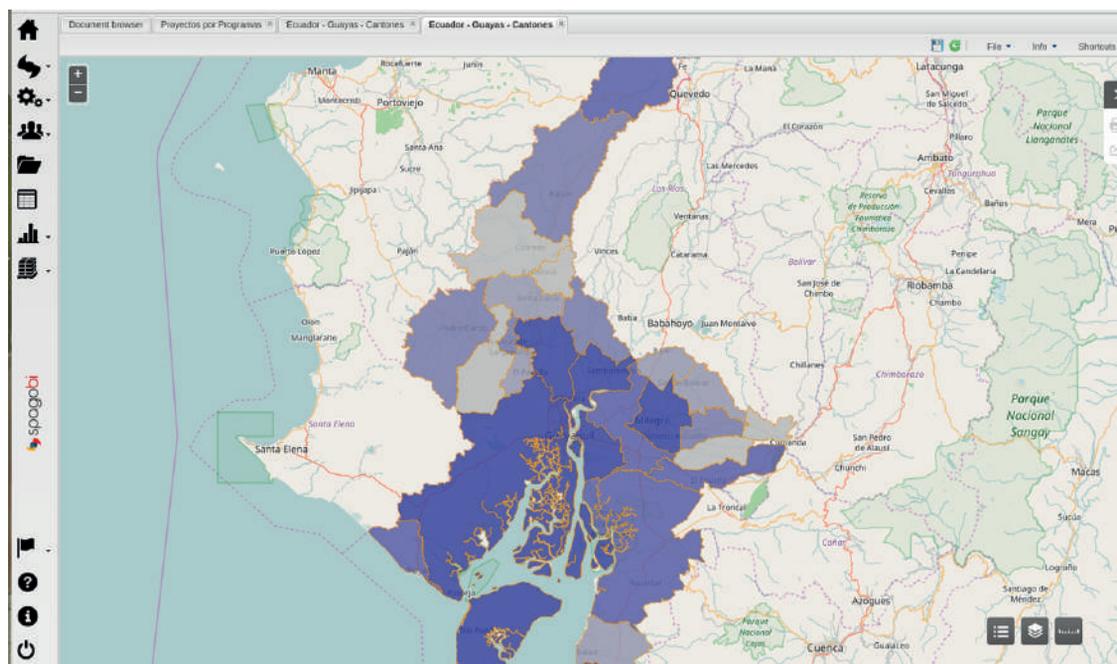


Figura 5. Mapa de la densidad poblacional de los cantones de la provincia del Guayas.

UNIVERSIDAD POLITÉCNICA SALESIANA
Dirección Técnica de Vinculación con la Sociedad
Sede Guayaquil

Parroquias por Cantón

Provincia: GUAYAS
Cantón: GUAYAQUIL

	1990	2001	2010
GUAYAQUIL	1,483,436	1,987,734	2,282,181
JUAN GÓMEZ RENDON (PROGR)	0	4,107	0
JUAN GÓMEZ RENDON (PROGRESO)	0	0	8,420
MORRO	0	2,619	3,882
POSORJA	0	18,224	23,957
PUNA	0	3,630	3,971
TENGUEL	0	6,954	8,865

Figura 6. Reporte de las parroquias del Cantón Guayaquil de la provincia del Guayas.

años 1990, 2001 y 2010, como se puede observar en la Figura 7.

Esta misma arquitectura se aplicó en otro caso de estudio más complejo relacionado con zonas de intervención en la ciudad de Guayaquil donde se pueden ejecutar proyectos de vinculación, publicado en (Pizarro *et al.*, 2018); en este artículo se presenta en detalle la arquitectura del Almacén de Datos Espacial utilizado.

La Arquitectura del ADE propuesto permite ejecutar de manera adecuada y automática el procedimiento ETL espacial y la visualización de los datos espaciales para los usuarios.

La automatización del procedimiento ETL es clave en un ADE, debido a que el esfuerzo no se concentra en la migración de los datos sino en el diseño de la visualización de los datos, el esfuerzo mayor en el procedimiento ETL solo fue

al inicio y una sola vez.

Para alguna institución que desee implementar la arquitectura de un ADE es factible económicamente, si se lo implementa tal y como se encuentra detallado en este artículo; o si desea brindar un mejor servicio al cliente se puede invertir en otro servidor, donde se alojaría sólo el *SpagoBI Server* y el *GeoServer*, esto sólo beneficiaría en un mejor rendimiento.

Los sistemas informáticos para ayuda en la toma de decisiones se extienden al ámbito geográfico, permitiendo a los usuarios finales conocer de mejor manera la distribución geográfica de sus procedimientos, otra forma de visualizar sus operaciones.

FOSS4G siempre es una alternativa viable para la implementación de proyectos que hagan uso de Sistemas de Información Geográficos, como se puede

verificar en este caso de estudio detallado.

En la base de datos *PostgreSQL* + *PostGIS*, donde se aloja el ADE, se puede implementar una extensión que permita el análisis de aprendizaje automático (*machine learning*); en este caso, con redes neuronales no supervisadas como las redes de Kohonen, debido a que se desconoce cómo se podrían clasificar los datos del INEC y visualizar esta clasificación en un mapa.

Un mapa geográfico permitiría verificar de mejor manera los patrones encontrados o por lo menos entender por qué se clasifican de una manera en particular.

Agradecimientos

Este trabajo de investigación se realizó gracias a la colaboración de la Dirección Técnica de Vinculación con la Sociedad de la Universidad Politécnica Salesiana, sede Guayaquil; y al Vicerrector de la Sede.

Referencias

- Astriani, W. y Trisminingsih, R. (2016). Extraction, Transformation, and Loading (ETL) module for hotspot spatial data warehouse using geokettle. *Procedia Environmental Sciences*, 33, 626-634.
- Berrahou, L., Lalande, N., Serrano, E., Molla, G., Berti-Équille, L., Bimonte, S., Bringay, S., Cernesson, F., Grac, C., Ienco, D., Le Ber, F., Teisseire, M. (2015). A quality-aware spatial data warehouse for querying hydroecological data. *Computers & Geosciences*, 85, 126-135.
- Bohorquez, J. E. (2003). Aproximación Metodológica de un Spatial Data Warehouse. Oracle DBA.
- González, D. B., y González, L. P. (2013). Spatial data warehouses and SOLAP using open-source tools. 2013 XXXIX *Latin American Computing Conference (CLEI)*. <https://doi.org/10.1109/CLEI.2013.6670660>.
- González, D. B., y González, L. P. (2014). *GeoCR: una aplicación web de análisis dinámico para el soporte de decisiones basadas en datos convencionales y espaciales*. Tesis de Grado. San José: Universidad de Costa Rica.
- Fiedler, J. (2011). *Harmonizace CAD a GIS dat na území Nečtin*. Tesis de Grado. Pilsen: Universidad de Bohemia Occidental. Chequia.
- Hasanah, G. U. y Trisminingsih, R. (2016). Multidimensional Analysis and Location Intelligence Application for Spatial Data Warehouse Hotspot in Indonesia using SpagoBI. En *IOP Conference Series: Earth and Environmental Science* (Vol. 31, p. 12011). IOP Publishing.
- Inmon, W. H. y Kelley, C. (1994). The 12 rules of data warehouse for a client/server world. *Data Management Review*, 4(5), 6-16.
- Oubennaceur, K. (2011). *Structuration d'une base de données sur l'agriculture en Rhône-Alpes*. Tesis de Maestría. Saint-Étienne: Université Jean Monnet. Francia.
- Pizarro, G., Jurado, V., Coque, Sh. (2018). Spatial Data Warehouse applied to Social Projects Management: Case Study. *Journal of Science and Research*, 3, 9.
- Pizarro, G. y Urvina, R. (2017). Diseño de un "Spatial Data Warehouse" y su incidencia en la generación de nuevos proyectos. En *Sistemas de Información en Ciencias de*

la Computación (pp. 15-65). Guayaquil:
Abya Yala.

Prunaire, F. X., Chartier, B., Coudert, M., y Jacolin,
Y. (2007). Spatial extension for Talend
(version 6.0.1). Disponible en:
<https://talend-spatial.github.io/>

Sale, H. B., Patil, D., Thube, S., y Student, B. E.
(2018). Crime Prevention with Data
Warehouse using OLAP through Business
Intelligence. *International Journal of
Engineering Science*, 8(2), 16017-16019.

Hasanah, G. U., y Trisminingsih, R. (2016, January).
Multidimensional analysis and location
intelligence application for spatial data
warehouse hotspot in indonesia using
spagobi. In *IOP Conference Series: Earth
and Environmental Science* (Vol. 31, No.
1, p. 012011). IOP Publishing.

