TEC | Tecnológico de Costa Rica

Escuela de Ingeniería en Computación
Bachillerato en Ingeniería en Computación
Sede Interuniversitaria de Alajuela

Prof. Alberto Shum Chan
Base de Datos II
Semestre I, 2024

Proyecto #3

Tema: extraer, transformar y cargar datos con SparkSQL

Entrega en el TecDigital:

- Adjuntar archivo .**zip** con el archivo en formato Jupyter notebook bien documentado que incluya los ejercicios y los datos utilizados.
- Forma de trabajo en grupos de 3 personas.
- Se agendará cita para revisión de la funcionalidad y dominio de lo implementado

Objetivo proyecto: Aplicar técnicas para extracción, transformación y carga a datos de la vida cotidiana generados por instituciones en Costa Rica, calcular estadísticas sobre los datos y visualizarlas por medio de Spark.

Objetivos específicos:

- Poner en práctica el conocimiento visto en clase sobre SparkSQL aplicado a datos de Costa Rica.
- Diseñar e implementar mecanismos que permitan analizar los datos por medio de estadísticas y visualización utilizando Spark.

Instrucciones:
Este proyecto busca que los estudiantes se expongan a la complejidad que implica obtener e integrar datos reales que provienen de múltiples fuentes. Para el ejercicio se utilizarán datos de criminalidad en Costa Rica combinados con datos socio-económicos asociados a los cantones del país. Los datos deberán ser preprocesados e integrados de manera que puedan ser utilizados para propósitos analíticos.

Los datos provienen de dos instituciones nacionales, a saber:

1. Instituto Nacional de Estadística y Censos de Costa Rica (INEC): El INEC es la institución encargada a nivel nacional de la generación y divulgación de datos estadísticos obtenidos por medio de censos, encuestas y otros estudios sobre demografía, economía y otros. Al igual que en el OIJ, los datos están disponibles por distrito. Los datos fueron generados por el INEC como resultado del censo realizado en el país en el año 2011. Los datos están disponibles en [1].
2. Organismo de Investigación Judicial (OIJ): El OIJ publica datos sobre criminalidad en Costa Rica que tienen como fuente las denuncias interpuestas directamente ante esta entidad nacional. Los datos recopilados por el OIJ están disponibles por provincia, cantón y

distrito (deben ser bajados como hoja electrónica para contar con el dato de distrito). Los datos están disponibles en [2].

El conjunto de datos de criminalidad del OIJ posee las siguientes columnas:

- Delito:Tipo de Delito
- SubDelito: Tipo de SubDelito
- Fecha: Fecha del Hecho
- Hora: Rango de 3 horas del Hecho
- Víctima: Descripción de la Víctima
- SubVictima: Descripción de la SubVíctima
- Edad: Grupo de Edad que pertenece la Víctima
- Género: Género de la Víctima
- Nacionalidad: Nacionalidad de la Víctima
- Provincia: Provincia del Lugar del Hecho
- Cantón: Cantón del Lugar del Hecho
- Distrito: Distrito del Lugar del Hecho

El segundo conjunto de datos contiene información sobre indicadores económicos, según provincia, cantón y distrito. Los datos fueron generados por el INEC como resultado del censo realizado en el país en el año 2011. El conjunto de datos posee las siguientes columnas:

- Provincia, Cantón y Distrito
- Población de 15 años y más
- Tasa neta de participación
- Tasa de ocupación
- Tasa de desempleo abierto
- Porcentaje de población económicamente inactiva
- Relación de dependencia económica

En este proyecto se implementarán funciones que permitan cargar, limpiar e integrar ambos conjuntos de datos de forma que puedan ser utilizados para el análisis de la cantidad y los tipos de delitos a nivel nacional. Investigaciones en análisis de datos asociados a delitos, han concluido que algunos factores, como son la educación, la pobreza y el desempleo pueden influir en el aumento o disminución de la tasa de delitos de una región (Suhong, Param, Parminder y Pooya, 2018) y es por eso que se requiere integrar ambas fuentes de datos.

Instrucciones:

- El tercer proyecto consiste en extraer, transformar y cargar datos que provienen de dos instituciones nacionales.
- Todos los procedimientos y funciones deben contener documentación interna completa que incluya, una descripción general, descripción de los parámetros de entrada, descripción de la salida y descripción de bloques relevantes.

- Todos los procesos de extracción y transformación **se deben realizar por medio de SparkSQL.**

A continuación, se describen con más detalle las operaciones que se deben realizar.

Actividades a realizar

1. Baje los conjuntos de datos, publicados por el OIJ y el INEC. Para ambos conjuntos de datos asegúrese de obtener los datos completos para todas las provincias, cantones y distritos.
2. **Utilizando únicamente la funcionalidad de Spark en Python**, integre los dos conjuntos de datos por medio del nombre del distrito. Para que la integración sea exitosa se debe primero preprocesar los datos de forma que el campo para la reunión o join en ambos conjuntos de datos, en este caso el distrito, coincida (documente muy bien todo el proceso). Es decir, usted debe asegúrese de que los nombres de los distritos de ambos conjuntos de datos se escriban igual para que la mayoría de los registros puedan ser tomados en cuenta en el ejercicio de análisis de datos. Para lograr esta meta debe programar las siguientes funciones:
    a. Una función que elimine los espacios en blanco de la columna distrito para usarse en ambos conjuntos de datos.
    b. Una función que convierta a minúsculas el contenido de la columna distrito para usarse en ambos conjuntos de datos.
    c. Una función que devuelva **la lista** de distritos del conjunto de datos del OIJ que no coinciden con ningún distrito del conjunto de datos del INEC.
    d. Una función que devuelva **la cantidad** de registros en el conjunto de datos del OIJ que no coinciden con ningún distrito del conjunto de datos del INEC.
    e. Edite, utilizando SparkSQL, los nombres de los distritos del INEC para que coincidan con algunos de los del OIJ.
3. **Guarde los datos limpios del INEC y del OIJ en una base de datos en PostgreSQL.**
4. **Visualización de datos. La visualización se realizará por medio del software Spark.**

Realice las siguientes visualizaciones (por medio de gráficos):

1. Compare la cantidad de delitos y la tasa de ocupación para los 10 distritos con más delitos en el país.
2. Grafique la cantidad de delitos por día de la semana para el distrito con más delitos.
3. Grafique la cantidad de delitos por tipo y por distrito. Es decir, para el distrito seleccionado se debe graficar la cantidad de delitos por tipo.
4. Grafique la cantidad de delitos por sexo para todo el conjunto de datos.
5. Proponga una visualización de su interés.

5. **Documentación: Se debe generar la siguiente documentación del proyecto.**

Contenido del documento:

- Descripción general del sistema.

- Descripción de las funciones. Todos los procedimientos y funciones deben contener documentación interna completa que incluya, una descripción general, descripción de parámetros de entrada, descripción de salida y descripción de bloques relevantes.
- Descripción de cada visualización.
- Conclusiones (al menos cuatro)
- Referencias

Referencias

[1] Instituto Nacional de Estadísticas y Censos (2011). Censo 2011: Indicadores económicos, según provincia, cantón y distrito. Recuperado de https://admin.inec.cr/sites/default/files/media/reempleocenso2011-22.xls_2.xls

[2] Organismo de Investigación Judicial (2018). Estadísticas policiales. Recuperado de https://sitiooij.poder-judicial.go.cr/index.php/apertura/transparencia/estadisticas-policiales

[3] Suhong, K., Param, J., Parminder, K. y Pooya, T. (2018). Crime Analysis Through Machine Learning. IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). Seleccionar del 1 de enero 2011 al 31 de diciembre 2011, que contemple todo el país, todas las categorías y todas las víctimas (47.481 registros).

Rúbrica

Rubro	Puntos
1. **Baje los conjuntos de datos, publicados por el OIJ y el INEC**. Los conjuntos de datos están completos, es decir, ambos conjuntos de datos incluyen todas las provincias, cantones y distritos del país.	2
2. Utilizando únicamente la funcionalidad de Spark en Python, realice lo siguiente:	
a. Programe y utilice una función que elimine los espacios en blanco de la columna distrito de ambos conjuntos de datos.	1
b. Programe y utilice una función que convierta a minúsculas el contenido de la columna distrito de ambos conjuntos de datos.	1
c. Programe una función que devuelva la lista de distritos del conjunto de datos del OIJ que no coinciden con ningún distrito del conjunto de datos del INEC.	1
d. Programe una función que devuelva la cantidad de registros en el conjunto de datos del OIJ que no coinciden con ningún distrito del conjunto de datos del INEC.	1
e. Edite, utilizando SparkSQL, los nombres de los cantones del INEC para que coincidan con algunos de los del OIJ.	2
f. Toda la funcionalidad fue incluida en un archivo Jupyter notebook	1
3. Guarde los datos limpios del INEC y del OIJ en una base de datos en PostgreSQL. 2	2
4. Visualización	
1. Compare la cantidad de delitos y la tasa de ocupación para los 10 distritos con más delitos en el país.	2
2. Grafique la cantidad de delitos por día para el distrito con más delitos.	2
3. Grafique la cantidad de delitos por tipo y por distrito.	2
4. Grafique la cantidad de delitos por sexo para todo el conjunto de datos.	2

5. Proponga e implemente una gráfica de su interés.	2
5. Documentación	
• Descripción general del sistema	1
• Descripción de las funciones	1
• Documentación de visualizaciones	2
• Conclusiones (al menos tres) respecto a los datos analizados (visualizaciones)	3
• Referencias	1
• Los documentos no tienen errores ortográficos	1
Total	30