Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Ciencias y Sistemas

Seminario de Sistemas 2 Sección A

Ing. Luis Alberto Vettorazzi España

Aux. Glen Abra-ham Calel Robledo

Práctica 1

Solución de Business Intelligence

OBJETIVOS

- El estudiante aprenda y utilice la herramienta Apache Spark para el manejo de Big Data.
- El estudiante aprenda y aplique transformaciones y acciones en Apache Spark.
- El estudiante implemente Apache Spark con Python para la presentación de resultados.
- El estudiante pueda realizar gráficas de datos que permitan su análisis.

DESCRIPCIÓN

La práctica tiene como objetivo principal que el estudiante implemente una solución en memoria de tal manera que pueda utilizar una herramienta para el procesamiento de Big Data y mostrar los resultados en gráficas.

Debido a la solución propuesta anteriormente y los resultados positivos, la empresa Market502 ha quedado satisfecha, por tal razón ahora son capaces de tomar decisiones acerca de sus ventas y compras, además de realizar ofertas en sus productos para tener una mayor demanda en los actuales y posibles compradores.

Teniendo en cuenta que ahora el proceso de almacenamiento de los cubos y procesamiento de los mismos no es una solución, los propietarios desean que se implementen nuevas tecnologías para el procesamiento de grandes cantidades de datos mucho más veloces, basadas totalmente en memoria y que de ser necesario pueden ser paralelizadas facilitando el crecimiento horizontal. Por esta razón se le pide una propuesta para análisis de datos en memoria.

IMPLEMENTACIÓN SUGERIDA

Debido a que la propuesta a implementar tiene que ser una solución de procesamiento en memoria, se le proporcionará únicamente información acerca de las ventas del negocio en un periodo dado para que realice las pruebas sobre estos datos. Se utilizará Apache Spark como el software que le permite realizar el procesamiento de grandes cantidades de datos en clusters de memoria y el lenguaje de programación Python, dada la cantidad de librerías disponibles para operaciones matemáticas que este lenguaje posee.

Por tal razón se le sugiere llevar a cabo los siguientes pasos:

1. Puede utilizar el sistema operativo que desee (se busca la mayor cantidad de software gratis para la reducción de costos en la empresa).
2. Instale Apache Spark preconstruido sobre Hadoop en su última versión.
3. Instale Python y los componentes necesarios para conectar con Spark y realizar las gráficas con la librería Plotly.
4. Realice scripts individuales para cada uno de los análisis solicitados.

REPORTES

Al ejecutarse el script deberá dar como resultado una gráfica que se abra en el navegador.
Los reportes solicitados son los siguientes:

1. **Archivo *suicide*:**
    a. Gráfica de barras que muestre el total de suicidios de las siguientes generaciones:
        Generation X
        Generation Z
        Boomers
        Silent
        G.I. Generation
        Millenials
    b. Gráfica de pie que muestre el total de suicidios por edades en Guatemala.

2. **Archivo *premier_league*:**
    a. Gráfica de pie que compare el resultado de cada partido del equipo Manchester United de la temporada 2017-2018 (ganador del partido: H para el equipo local, A para el equipo visitante, D para el empate).
    b. Gráfica de barras que muestre el total de goles fuera de casa (gol de visita) por club de la temporada 2012-2013.

3. **Archivo *mass_shootings*:**
    a. Gráfica de barra que reporte el año con más tiroteos.
    b. Gráfica de barra que reporte el año con más muertes.
    c. Gráfica a su elección, de muertes, heridos y total de víctimas del año 2015.

RESTRICCIONES

- Utilizar Apache Spark como el software de procesamiento de datos en memoria.
- Se debe utilizar Python.
- Únicamente se puede utilizar el módulo SQLContext para la lectura de archivos (posteriormente deberá realizar la conversión del DataFrame a RDD).
- Para realizar los reportes únicamente pueden utilizar Transformaciones y Acciones sobre RDD.
- Para cada inciso se debe entregar scripts individuales.
- Los reportes libres pueden utilizar la gráfica que deseen, siempre y cuando sean entendibles.

CONSIDERACIONES

- La entrega es individual.
- Fecha de entrega: entregar scripts y documentación con screenshots de los reportes en la publicación de Google Classroom identificado como [SS2]Practica1_carne.zip **identificado como [SS2]Practica1_carnet.zip** el día viernes 24 de abril.
- Entregas tarde se califican con un 50% de penalización.
- Copias detectadas tendrán nota 0 y reporte a la escuela.