Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Ciencias y Sistemas

Seminario de Sistemas 2 Sección A

Ing. William Estuardo Escobar Argueta

Aux. Glen Abra-ham Calel Robledo

Práctica 1

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(Opcional).
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 *video_games_sales*:**
    a. Gráfica de barras que muestre el total de ventas globales de las siguientes categorías:
        Action
        Sports
        Fighting
        Shooter
        Racing
        Adventure
        Strategy
    b. Gráfica de pie que muestre el total de géneros publicados por la plataforma Nintendo.
    c. Gráfica de barras que muestre el top 5 de plataformas con más lanzamientos.

2. **Archivo *sales*:**
    a. Gráfica de pie que muestre una comparación de los ingresos de todas las regiones (Centro América, Europa, Asia, etc).
    b. Gráfica a su elección, que reporte cuál es el año con más unidades vendidas en Guatemala.
    c. Gráfica a su elección, de ganancias (ingresos - costos), ingresos y costos del 2010, de las ventas en línea.
3. **Archivo *police_killings*:**
    a. Gráfica de barra que muestre el top 3 de las razas de las víctimas.
    b. Gráfica de barra que muestre el top 5 de años con más incidentes.

DOCUMENTACIÓN

- Requisitos del sistema.
- Versiones de las herramientas que utilizó.
- Descripción de las transformaciones y acciones que utilizó.
- Captura de pantalla con los resultados obtenidos.

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 numeración se debe entregar scripts individuales.
- Los reportes a su elección pueden utilizar la gráfica que deseen, siempre y cuando sean entendibles.

CONSIDERACIONES

- La entrega es individual.
- Fecha de entrega: entregar scripts y documentación en la plataforma UEDI identificado como **[SS2]Practica1_carnet.zip** el día lunes 22 de junio.
- Entregas tarde se califican con un 50% de penalización.
- Copias detectadas tendrán nota 0 y reporte a la escuela.
- La calificación se realizará sobre lo que se envía.