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. Breynner Miguel Cortez Sic

Práctica 2
Apache Spark - Python

OBJETIVOS:

- Que el estudiante aprenda cómo trabaja la herramienta Apache Spark para el manejo de Big Data.
- Que el estudiante experimente el análisis con datos estructurados con las herramientas de Apache Spark y Python para la presentación de resultados.
- Que el estudiante aprenda y sepa cuándo aplicar transformaciones y acciones en Apache Spark.
- Que el estudiante experimente el análisis de datos y llevar este a una forma gráfica.

DESCRIPCIÓN:

Debido a la buena implementación que ha tenido con lo solicitado por GuateFood, la empresa ha quedado satisfecha. Debido a esto lo ha contactado nuevamente pero esta vez para el procesamiento de grandes cantidades de datos que más adelante se detallarán.

Teniendo en cuenta la cantidad de datos y la estructura de estos, la implementación de cubos y el procesamiento de estos no es una solución, por esto se desea implementar nuevas tecnologías para el procesamiento de estas grandes cantidades de datos de manera veloz, basadas totalmente en memoria y garantizar su buen funcionamiento y la aceptación a crecimiento de estos datos. Por este motivo se le solicita una propuesta para el análisis de datos en memoria.

IMPLEMENTACIÓN SUGERIDA

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.

- Se podrá utilizar el sistema operativo que usted desee.
- Instale Apache Spark preconstruido sobre Hadoop en su última versión.
- Instale Python y los componentes que crea necesarios para conectar con Spark, entre ellos la librería Plotly para la realizar las gráficas.
- Con la ayuda del software Anaconda, inicie el IDE Jupyter Notebook y realice los reportes.

Como apoyo se le proporciona algunos links de instalación de las herramientas:

- Opción 1
- Opción 2
- Link para Plotly

FLUJO DE DATOS

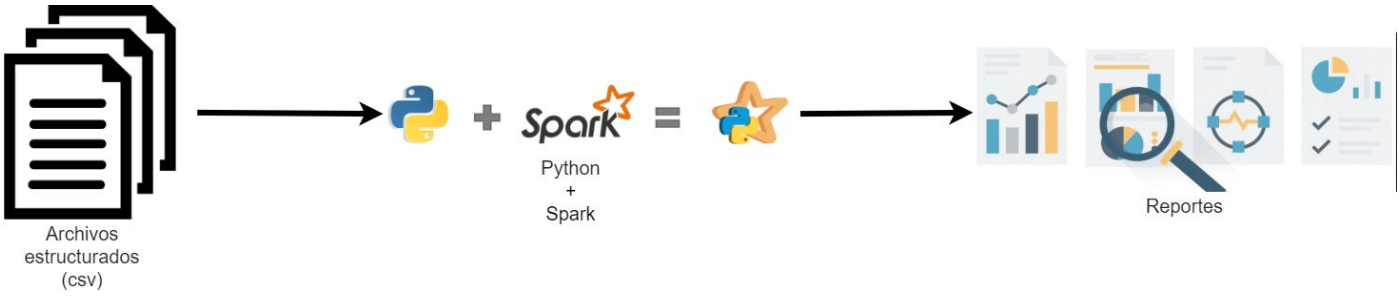

Diagrama 1.

REPORTES

Se tendrán 3 archivos en los cuales se detallan los siguientes reportes para su presentación:

Archivo1 ***GuatemalaExportsTo***:

1. Gráfica de barras que muestre el país con el mayor valor por exportaciones.
2. Gráfica de pie de los 5 países con menos valor por exportaciones.

Archivo2 ***TraficoAereoGt***:

1. Gráfica de barra con el total de aterrizajes por Aeropuerto.
2. Gráfica de pie con los 3 meses con mayor número de pasajeros de salida.

Archivo3 ***Covid19***:

1. Gráfica de barras con el número total de casos de COVID-19 en
    - Cuba
    - France
    - Canada
    - Singapore
    - South_Korea
2. Gráfica de pie con los 5 meses con menos casos de COVID.
3. Gráfica a su elección del total de casos y muertes en Guatemala en el mes de Agosto.

RESTRICCIONES

- Se debe utilizar Apache Spark como software de procesamiento de datos en memoria.
- Se debe utilizar Python.
- El módulo SQLContext será permitido sólo para lectura de archivos (posteriormente deberá realizar la conversión del DataFrame a RDD).
- Para la realización de reportes se permite únicamente ***Transformaciones*** y ***acciones*** sobre RDD.
- Debe entregar scripts individuales por reporte con el siguiente formato: ***R#_A#.txt*** siento *R#* el número de reporte y *A#* el número de archivo.

ENTREGABLES

- Scripts individuales por cada reporte con el formato ***R#_A#.txt***.
- Documentación con las imágenes de cada reporte.

CONSIDERACIONES

- La entrega es individual. **No habrá prórroga.**
- Todas las dudas con respecto a esta Fase deberán ser planteadas en los foros creados en la plataforma UEDI o en caso muy especial al correo breynner0195@gmail.com
- Enviar el proyecto vía UEDI en un zip con el nombre: **[SS2]P2_carne.zip el día viernes 30 de octubre de 2020 a las 23:59 horas.**
- Entregas tarde no se calificarán.
- De encontrar copias se tendrá una nota de 0 y el reporte a la escuela de sistemas.