Curso: Big Data Analytics Avanzado

Tema de la sesión: Aplicaciones de Big Data Analytics con Python

lunes, 25 de marzo de 2024

Carrera: Big Data y Ciencia de Datos
Código: C28A

Mg. Ing. Ernesto Cancho Rodríguez, MBA de la Universidad George Washington (EE.UU.)

Catedrático UNMSM

Capacidades Terminales

- Entender los fundamentos de Big Data Analytics y su importancia en el procesamiento de grandes volúmenes de datos con Python.
- Conocer y aplicar las bibliotecas de Python como Pandas, NumPy, y PySpark para el análisis y manejo de datos a gran escala.

Contenido

- Big Data Analytics con Python
- Caso de Laboratorio: PySpark en Colab
- Análisis de PySpark
- Desarrollo de Aplicaciones de Big Data Analytics con Python

Big Data Analytics con Python (1)

Big Data Analytics se refiere al proceso de analizar y extraer información de grandes volúmenes de datos, que no podrían ser procesados de manera convencional.

Python se ha convertido en una herramienta esencial en este campo debido a su simplicidad y a la poderosa suite de bibliotecas disponibles para el análisis de datos, como **Pandas**, **NumPy**, y **PySpark**.

Estas herramientas permiten manejar, analizar y visualizar grandes conjuntos de datos de manera eficiente.

Big Data Analytics con Python (2)

Ejemplo:

Un ejemplo de **Big Data Analytics** es el análisis de datos de redes sociales para entender las tendencias de consumo de los usuarios.

Utilizando **Python** y sus bibliotecas, podemos recopilar datos, procesarlos y aplicar modelos de machine learning para predecir comportamientos futuros.

Caso de Laboratorio:
PySpark en Colab, Analítica de Bolsa

Este notebook de **Colab** proporciona un ejemplo práctico del uso de **PySpark** para el análisis y visualización de datos financieros de **NVIDIA**.

El documento comienza con la configuración del entorno de Colab para soportar **PySpark**, seguido de la carga de datos financieros y su posterior análisis y visualización utilizando **DataFrames** de **PySpark**.

Caso de Laboratorio: PySpark en Colab (2)

Código y Datos:

El caso detalla cómo instalar **PySpark** en **Colab**, cargar datos desde un archivo CSV, y realizar operaciones básicas de análisis de datos como filtrados y agregaciones. Además, se muestra cómo visualizar los datos financieros, proporcionando insights valiosos sobre el rendimiento de la empresa.

Caso de Laboratorio: PySpark en Colab (3)

Para entender el caso, es crucial familiarizarse con conceptos clave de **PySpark** como **RDDs**, **DataFrames**, **Spark SQL**, y las operaciones de transformación y acción.

Los **RDDs** son colecciones distribuidas de datos, sobre las cuales se pueden realizar operaciones en paralelo.

Los **DataFrames** ofrecen una interfaz más alta, permitiendo operaciones de datos similares a las bases de datos.

Análisis de PySpark (1)

PySpark es la interfaz de **Python** para **Apache Spark**, un motor de análisis y procesamiento de Big Data que permite el procesamiento de datos en clústeres de manera distribuida y paralela.

PySpark facilita el trabajo con grandes volúmenes de datos a través de **RDDs (Resilient Distributed Datasets)** y **DataFrames**, abstracciones que permiten manipular los datos de forma distribuida y optimizada.

Análisis de PySpark (2)

Ejemplo: Un caso práctico de uso de **PySpark** es la optimización de rutas de logística basada en datos históricos de entregas.

Imaginemos un sistema con **PySpark**, con el cual se pueden analizar millones de registros de entregas para encontrar patrones y optimizar las rutas futuras.

Desarrollo de Aplicaciones de Big Data Analytics con Python (1)

El desarrollo de aplicaciones de **Big Data Analytics** implica varias etapas, incluyendo la recopilación de datos, su limpieza y preprocesamiento, el análisis exploratorio, la aplicación de modelos estadísticos o de machine learning, y finalmente, la visualización y comunicación de los resultados.

Python provee un ecosistema rico de bibliotecas y herramientas para cada una de estas etapas, haciendo el proceso eficiente y accesible.

Desarrollo de Aplicaciones de Big Data Analytics con Python (2)

Ejemplo: Un ejemplo sería el desarrollo de un sistema de recomendación para una plataforma de streaming utilizando **Python**.

A través de la recopilación de datos de interacciones de los usuarios, se pueden aplicar algoritmos de filtrado colaborativo con **PySpark** para recomendar contenido relevante a cada usuario.


Tecsup®
TECNOLOGÍA
CON SENTIDO