Tarea online BDA03.

Título de la tarea: Almacenamiento y procesamiento en Hadoop.

Curso de especialización y módulo: Especialización en Inteligencia Artificial y Big Data - Big Data Aplicado

¿Qué contenidos o resultados de aprendizaje trabajaremos?

Resultados de aprendizaje

- **RA1.** Gestiona soluciones a problemas propuestos, utilizando sistemas de almacenamiento y herramientas asociadas al centro de datos.
- **RA2.**Gestiona sistemas de almacenamiento y el amplio ecosistema alrededor de ellos facilitando el procesamiento de grandes cantidades de datos sin fallos y de forma rápida.
- **RA3.**Valida las técnicas de Big Data para transformar una gran cantidad de datos en información significativa, facilitando la toma de decisiones de negocios.

Contenidos

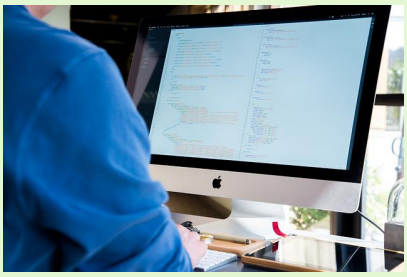
1.- Introducción al ecosistema Hadoop.
2.- Componentes de acceso y procesamiento de datos.
    2.1.- Apache Pig.
    2.2.- Apache Hive.
        2.2.1.- Conceptos generales.
        2.2.2.- Arquitectura.
        2.2.3.- HQL.
    2.3.- Apache Impala.
    2.4.- Apache HBase.
    2.5.- Apache Phoenix.
    2.6.- Apache Spark.
        2.6.1.- Arquitectura y componentes.
        2.6.2.- Detalle de los componentes de Apache Spark.
        2.6.3.- Ventajas y desventajas.
3.- Componentes de ingesta de datos y flujos de trabajo.
    3.1.- Apache Sqoop.
    3.2.- Apache Flume.
    3.3.- Apache Oozie.
4.- Interfaces y herramientas de trabajo.
    4.1.- Hue.
    4.2.- Apache Zeppelin.
    4.3.- Apache Ambari y Cloudera Manager.
5.- Procesamiento en streaming: Apache Spark (Structured Streaming), Apache Flink y Apache Storm.
6.- Guía práctica Apache Hive y Apache Pig.

Siguiente »

1.- Descripción de la tarea.

Consultas en Apache Hive y en Apache Pig

En esta práctica aprenderás a utilizar las dos herramientas más populares para manipulación de datos en Apache Hadoop: Hive y Pig. Utilizarás el mismo *dataset* que se utilizó en la guía práctica para resolver la misma consulta con ambas herramientas para que aprendas sus diferencias y puedas comparar su uso. Para resolver las consultas deberás consultar la documentación de uso de HQL (Hive) y Pig Latin (Pig).



Pexels (Dominio público)

¿Qué te pedimos que hagas?

Descarga el enlace al fichero zip inferior, descomprímelo y sigue las instrucciones que se dieron en la guía práctica. En resumen lo que hay que hacer navegar al servidor de Jupyter, abrir el libro de Jupyter (el que tiene extensión "ipynb") y seguir las instrucciones de realización que se dan en él.

Para entregar debes subir un fichero comprimido que contenga el fichero "ipynb" con las celdas ejecutadas.

- Tarea_BDA03.zip (Ventana nueva)

NOTA IMPORTANTE

Para todas las preguntas es necesario entregar un documento en el que aparezca la pregunta junto con tu respuesta razonada. Incluye cualquier diagrama que consideres interesante para explicar la solución que has desarrollado.

2.- Información de interés.

Recursos necesarios y recomendaciones

Recursos necesarios

- Conexión a Internet para consultar la información o investigar sobre las cuestiones planteadas.
- Navegador web.
- Equipo informático con procesador equivalente a un i5 o superior y un mínimo de 8GB de RAM.
- Recomendable SO Linux. No te preocupes si no lo tienes instalado, también puedes hacer la práctica con Windows o MAC.

Recomendaciones

- Antes de abordar la tarea:
  - lee con detenimiento la unidad, consulta los enlaces para saber más, revisa la guía práctica y realiza paso a paso lo que allí se indica.
  - Si tienes problemas, escribe en el foro.
  - Realiza el examen online de la unidad, y consulta nuevamente las dudas que te surjan. Solo cuando lo tengas todo claro, debes abordar la realización de la tarea.
- No olvides crear un "zip" que contenga el archivo "ipynb" con las celdas ejecutadas y las imágenes que hayas utilizado.

Indicaciones de entrega

Una vez realizada la tarea, el envío se realizará a través de la plataforma. El archivo se nombrará siguiendo las siguientes pautas:

Apellido1_Apellido2_Nombre_BDA03_Tarea

Asegúrate que el nombre no contenga la letra ñ, tildes ni caracteres especiales extraños. Así por ejemplo la alumna **Begoña Sánchez Mañas para la tercera unidad del MP de BDA**, debería nombrar esta tarea como...

sanchez_manas_begona_BDA03_Tarea

3.- Evaluación de la tarea.

Criterios de evaluación implicados

Criterios de evaluación RA1

- a) Se ha caracterizado el proceso de diseño y construcción de soluciones en sistemas de almacenamiento de datos.
- b) Se han determinado los procedimientos y mecanismos para la ingestión de datos.
- c) Se ha determinado el formato de datos adecuado para el almacenamiento.
- d) Se han procesado los datos almacenados,
- e) Se han presentado los resultados y las soluciones al cliente final en una forma fácil de interpretar.

Criterios de evaluación RA2

- a) Se ha determinado la importancia de los sistemas de almacenamiento para depositar y procesar grandes cantidades de cualquier tipo de datos rápidamente.
- b) Se ha comprobado el poder de procesamiento de su modelo de computación distribuida.
- c) Se ha probado la tolerancia a fallos de los sistemas.
- d) Se ha determinado que se pueden almacenar tantos datos como se desee y decidir cómo utilizarlos más tarde.
- e) Se ha visualizado que el sistema puede crecer fácilmente añadiendo módulos.

Criterios de evaluación RA3

- a) Se ha realizado la limpieza y transformación de datos en base a los objetivos predeterminados.
- b) Se ha comprobado el poder de procesamiento de su modelo de computación distribuida.
- c) Se han conjugado dentro de un modelo de empresa datos de clientes, financieros de ventas, de productos, de marketing, de redes sociales, de la competencia, entre otros, para extraer un análisis valioso y efectivo para el negocio.

¿Cómo valoramos y puntuamos tu tarea?

Rúbrica de la tarea	
Pregunta 1	2 puntos
Pregunta 2	2 puntos
Pregunta 3	2 puntos
Pregunta 4	2 puntos
Pregunta 5	2 puntos
Redacción clara y correcta, sin errores ortográficos	**Se resta 0,1 puntos por cada error ortográfico o expresiones incorrectas.**

« Anterior