24-38/000306 IFCD104

Inteligencia Artificial y Big Data aplicado al ámbito biosanitario

ESCALA DE CALIFICACIÓN E_1

INSTRUCCIONES PARA EL ALUMNADO

MÓDULO:	Módulo 3
ACTIVIDAD:	**E_1**: Desarrollo de una aplicación de Big Data
FECHA:	31/03/2025
DURACIÓN:	6 horas
NOMBRE Y APELLIDOS	

Descripción de la práctica

En esta actividad, configurarás una máquina virtual con Ubuntu Server, instalarás y configurarás Hadoop en modo pseudo-distribuido, y ejecutarás operaciones de procesamiento de datos utilizando HDFS y MapReduce. El objetivo es adquirir experiencia práctica en la implementación de un entorno de Big Data local y comprender el flujo de trabajo en Hadoop.

Instrucciones específicas

Parte 1: Preparación del Entorno

Configuración de la Máquina Virtual

1. Usa VirtualBox o VMware para crear una máquina virtual con Ubuntu Server.
2. Asigna al menos 4GB de RAM y 2 núcleos de CPU.
3. Configura una interfaz de red en modo "Bridge" o "NAT" con acceso a Internet.
4. Instala Hadoop y configúralo en modo pseudo-distribuido. Asegúrate de que los servicios de Hadoop estén en ejecución.

Parte 2: Procesamiento de Datos con Hadoop y Herramientas Relacionadas

Cargar y Procesar Datos en HDFS

1. Descarga un dataset grande (por ejemplo, logs de acceso a servidores o un dataset de Open Data).

- Copia los datos a HDFS ejecutando los comandos necesarios.

Ejecutar un Job de MapReduce

1. Ejecuta el programa de ejemplo wordcount sobre los datos almacenados en HDFS:

2. Verifica los resultados obtenidos ejecutando..

Entrega y Evaluación

- Captura de pantalla de la configuración de la máquina virtual.
- Comprobación de la instalación y ejecución de Hadoop.
- Evidencia de los datos cargados en HDFS y ejecución del Job de MapReduce.
- Análisis breve sobre los resultados obtenidos y posibles mejoras en el procesamiento.

Equipo y material

Hardware

- Computadora con al menos 8GB de RAM, procesador de 4 núcleos y 50GB de almacenamiento libre.
- VirtualBox o VMware instalado para crear una máquina virtual.

Software

- Ubuntu Server instalado en la máquina virtual.
- Hadoop configurado en modo pseudo-distribuido.
- Java (JDK 8 o superior) para compatibilidad con Hadoop.
- SSH y herramientas de red para la administración remota.

Datos y Herramientas

- Dataset grande, como logs de acceso a servidores o datos de Open Data.
- HDFS para almacenamiento distribuido.
- MapReduce para procesamiento de datos.

Conectividad

- Configuración de red en modo Bridge o NAT con acceso a Internet.