IES de Teis
Avda. de Galicia, 101
36216 – Vigo

Tfno: 886 12 04 64
e-mail: ies.teis@edu.xunta.es
http://www.iesteis.es

36018173. CSIFC90. MP5075. Big Data Aplicado. 2023-2024. (Grupo A)

Parcial JUNIO24

Nombre completo y DNI

Leer antes de empezar.

- La duración del examen es de 2 horas con 5 minutos de cortesía.
- Al acabar el examen debes firmar en la hoja de entrega.
- Ni se permitirá la firma en la hoja de entrega ni se recogerán exámenes fuera de tiempo.
- Durante esta prueba escrita puedes utilizar libremente todo tu material escrito y digital.
- No estará permitido comunicarse con el resto de alumnado por ningún medio.
- La prueba escrita debe permanecer inmutable por lo que es obligatorio usar bolígrafo permanente y no escribir en lápiz, usar corrector o cualquier otro método que pueda poner en duda una alteración posterior.
- Es necesario usar lenguaje técnico.
- Se puede solicitar hojas extra si se necesitan.
- El espacio disponible es suficiente para contestar cada pregunta.
- Las preguntas tipo test tienen una única respuesta válida. Preguntas incorrectas puntúan en negativo.

1.- Explica con tus palabras cual es la diferencia entre una fase “shuffle” y una fase “combiner”

RESPUESTA: La etapa “shuffle” es obligatoria mientras que la etapa “combiner” es opcional. En la etapa “shuffle” se ordena y particiona. En la etapa “combiner” se hace una agregación previa al “shuffle” consistente en aplicar alguna operación sobre los valores de las mismas claves con el fin de reducir la cantidad de pares que pasan a la siguiente fase.

La etapa “shuffle” mantiene el mismo número de parejas clave-valor que recibe como entrada. La etapa “combiner” no escribe en disco tiene como objetivo reducir el número de pares clave-valor.

2.- Explica con tus palabras que función tiene el “Application Master”

RESPUESTA; Es un aplicación que se ejecuta en el primer contenedor de un trabajo distribuido en YARN y se encarga de solicitar nuevos contenedores así como gestionar y monitorizar los que ya tiene asignados.

3.- Te dan la opción de crear un clúster Hadoop enfocado principalmente al procesamiento distribuido. Como opción A dispones de 1000 Raspberry Pi 3 (4 núcleos y 1 giga RAM). Como opción B dispones de 200 PCs (4 núcleos y 8 gigas de RAM). Suponiendo el resto de propiedades del clúster iguales, ¿qué opción recomendarías? Justifica la respuesta.

RESPUESTA: el valor mínimo por defecto de los contenedores es de 1 procesador (no habría problema porque cada nodo tiene 4) y 1 giga de memoria. Las rapsberrys no tendrían ese giga puesto que parte estaría ocupado por el sistema operativo. Tendrían que usar memoria virtual muy lenta. Aparentemente tendríamos mucha más memoria con la que crear contenedores y por tanto ejecutar distribuido pero la realidad es que escasamente cada nodo podría gestionar 1 contenedor. En la opción B podríamos tener sin problemas 3 contendores reales con su giga y procesador dedicado. Opción preferida la B. Quien justifique que se pueden crear contenedores más pequeños modificando propiedades también OK.

4.- Explica con tus propias palabras qué es un “Node Manager” y qué funciones tiene.

RESPUESTA: Es un componente software que se ejecuta en cada nodo trabajador del clúster Hadoop. Tiene dos funciones importantes. La primera, mantener informado al “Resource Manager” en cuanto a carga de trabajo y disponibilidad de recursos. La segunda, la creación y eliminación de contenedores dentro de su nodo así como la planificación de ejecución entre ellos.

5.- Explica con tus palabras la diferencia que hay entre YARN y los motores de procesamiento distrinuido.

RESPUESTA: YARN es la capa de gestión de recursos del clúster, se encarga de proveer y gestionar unidades atómicas de procesamiento distribuido en los nodos atendiendo a la carga de los nodos y la localidad de los datos en la capa HDFS. El motor de procesamiento se encarga de como procesar lógicamente los datos como por ejemplo MapReduce que lo convierta a pares clave-valor y después los agrega.

6.- Indica si el siguiente código pertenece a un mapper o a un reducer y justifica tu respuesta.

```python
#!/usr/bin/env python3
import sys
import json

def examen():
    for line in sys.stdin:

        record = json.loads(line.strip())
        dni = record.get('dni')
        altura = record.get('altura')

        if dni and altura:
            print(f"{dni}\t{altura}")

```

RESPUESTA: es un mapper porque toma valores de un json que puede indicar que es un registro de una base de datos o un log, en cualquier caso un archivo con datos en bruto. Después los recorre para filtrar solo DNI como clave y altura como valor.

7.- Indica si el siguiente psudocódigo pertenece a un mapper o un reducer y justifica tu respuesta.

```
INICIO

# Importar librerías necesarias
importar sys
importar pyhive.hive como hive

# Definir función
```

IES de Teis
Avda. de Galicia, 101
36216 – Vigo

Tfno: 886 12 04 64
e-mail: ies.teis@edu.xunta.es
http://www.iesteis.es


```
función examen():
    # Conectarse a Hive
    conexión = hive.connect('nombre_servidor', puerto, 'usuario', 'contraseña')
    cursor = conexión.cursor()

    # Definir la consulta SQL
    consulta_sql = """
    SELECT *
    FROM nombre_tabla
    WHERE ciudad = 'Vigo'
    """

    # Ejecutar la consulta SQL
    cursor.execute(consulta_sql)

    # Iterar sobre los resultados de la consulta
    para cada (dni, edad) en cursor.fetchall():
        # Emitir DNI y edad
        imprimir(dni + "\t" + edad)

    # Cerrar la conexión
    cursor.close()
    conexión.close()

# Llamar a la función principal
si __nombre__ == "__main__":
    examen()

FIN
```

RESPUESTA: aunque no sería un mapper MapReduce adecuado (los resultados de la consulta deberían estar en HDFS y no ejecutarse dentro del mapper) la idea es exactamente la de un mapper. Toma valores en bruto de una base de datos y los filtra emitiendo solo dni como clave y edad como valor.

8.- Indica si el siguiente código pertenece a un mapper o a un reducer y justifica tu respuesta.

```python
#!/usr/bin/env python3
import sys

def examen():
    conjunto_macs = set()

    for line in sys.stdin:
        mac_address, _ = line.strip().split("\t")
        if mac_address:
            conjunto_macs.add(mac_address)

    for mac in conjunto_macs:
        print(mac, "\t", "1")
```

RESPUESTA: es un reducer porque lee parejas de clave valor de las que solo guarda la mac e ignora el valor. Y porque después la va acumulando en un conjunto donde no hay repetición de elementos. Por último imprime los valores del conjunto. Este reducer agrega todas las macs que le ha llegado e imprime solo una aparición independientemente del número de apariciones reales que han enviado los mappers.

9.- ¿Cuál es la función principal de YARN en Hadoop?

- ☐ Proveer almacenamiento distribuido
- ☐ **Administrar recursos y tareas en un clúster de Hadoop**
- ☐ Realizar consultas SQK sobre datos distribuido
- ☐ Proveer una interfaz gráfica para gestionar Hadoop

10.- En el motor MapReduce, ¿qué fase se encarga de procesar las parejas clave-valor y producir una salida final?

- ☐ **Los reducers**
- ☐ Los Combiner
- ☐ Shuffle y sort
- ☐ Los mappers

11.- ¿Qué componente de YARN es responsable de iniciar y monitorizar las aplicaciones?

- ☐ **ResourceManager**
- ☐ NodeManager
- ☐ DataNode
- ☐ NameNode

12.- ¿Qué almacena el Hive Metastore en Apache Hive?

- ☐ Los datos de usuario

- ☐ La configuración del clúster Hadoop
- ☐ **La información de los esquemas y metadatos de las tablas Hive**
- ☐ Los resultados de las consultas SQL

13.- ¿Cuál es la principal función de Apache Hive en el ecosistema Hadoop?

- ☐ Almacenamiento distribuido de los datos
- ☐ **Ejecución de consultas SQL sobre grandes volúmenes de datos**
- ☐ Procesamiento de flujos de datos en tiempo real
- ☐ Gestión de la seguridad y autenticación de usuarios.

14.- Explica con tus propias palabras qué es Hadoop-Streaming y para qué se utiliza.

RESPUESTA: es una utilidad de Hadoop que permite implementar mappers y reducers en cualquier lenguaje de programación que pueda leer y escribir desde consola. Funciona redirigiendo la entrada y salida de los mappers y reducers desde su hilo de ejecución en java hasta la interfaz estándar donde lo recogerá. Se utiliza para disponer de más libertad a la hora de seleccionar el lenguaje de programación con el que implementar mappers y reducers.

15.- Explica con tus propias palabras qué hace este comando detallando cada uno de sus parámetros.

```
yarn jar /ruta/a/hadoop-streaming.jar
-input /ruta/a/carpeta1
-output /ruta/a/carpeta2
-file archivo1.txt
-mapper /ruta/a/archivo2.py
-file archivo2.py
-reducer /ruta/a/archivo3.sh
-file archivo3.sh
```

RESPUESTA: lanza un trabajo mapreduce mediante hadoop-streaming porque el mapper y el reducer está escrito en otros lenguajes de programación distinto de Java. La primera línea lanza hadoop-streaming para ejecutar el motor mapreduce desviando los mappers y los reducers. Input especifica la ruta HDFS donde están los datos de entrada. Output indica la carpeta HDFS donde se dejará los datos finales ya procesados. Los files indica archivos que se copiarán en los contenedores. Se enviará archivo1.txt que puede contener datos comunes a todos los contenedores, el propio mapper y reducer. El parámetro mapper especifica la ruta local del archivo que se ejecutará como mapper,

en este caso es un Python. El parámetro reducer especifica la ruta local del archivo con el código del reducer, en este caso es un script shell.

16.- Explica con tus palabras qué ventajas aporta la librería Python MRJob relacionándolo con los problemas que te has podido encontrar al realizar las prácticas de YARN.

RESPUESTA: fundamentalmente 2. La ejecución en local permite no tener que modificar el código fuente para que lea de ficheros en lugar de la consola y así es más sencillo probar y depurar. La otra ventaja es que MRJob gestiona la creación y eliminación de las carpetas con los resultados finales permitiendo ejecutar consecutivamente trabajos sin problemas.

17.- Explica con tus propias palabras cómo se configura la cantidad máxima de memoria que un contenedor de MapReduce puede usar.

REPUESTA: Al ser específico de MapReduce hay que modificar la propiedad mapreduce.map.memory.mb y la propiedad mapreduce-reduce.memory.mb en el archivo mapred-site.xml. Lo importante es tener claro el archivo donde mirar y aproximadamente el nombre de las propiedades.

18.- Explica con tus propias palabras cómo configurar la cantidad de núcleos asignados a los contenedores.

RESPUESTA: En este caso hay que editar el archivo yarn-site.xml con la propiedad yarn.nodemanager.resource.cpu-vcores.

19.- Explica con tus palabras cómo configurar la ruta a la carpeta donde Hive guardará los datos de las tablas.

RESPUESTA: en el archivo hive-site.xml hay que configurar la propiedad hive.metastore.warehouse.dir indicando la ruta HDFS donde Hive guardará los datos se las tablas.

20.- En Apache Hive, ¿qué diferencia hay entre una tabla interna y una externa?

RESPUESTA: los datos de una tabla interna se guardan dentro del warehouse de Hive por lo que si se borran con un DROP se pierden definitivamente. Los datos de una tabla externa son referenciados desde otra ruta HDFS por lo que al hacer un DROP de la tabla no se pierden los originales.