Prepara tu examen de BDA

Introducción

Cada centro, cada año y cada docente, puede plantear al alumnado un modelo de examen concreto que, a su criterio, pueda servir como una correcta evaluación del módulo.

Para ayudar a preparar las evaluaciones, he pensado que podría ser de ayuda crear un archivo único para cada módulo, que pueda crecer cada año con el feedback y apoyo de la comunidad, con cuestionarios de todo tipo, con solucionario o solo los enunciados, pues la intención primera es poder ofrecer una idea de lo que podemos encontrarnos a la hora de una evaluación, poder aprender con ello, y no algo que una persona acabe memorizando, y esperando, sin comprender ni ahondar en la materia, que aparezca mágicamente en el examen.

Este documento, por tanto, no pretende ser una guía única y veraz de exámenes pasados o futuros, pero si una fuente de información sobre la que basar vuestros estudios.

Posibles modelos.

Modelo 1.

1.- ¿Cómo se llama el sistema de almacenamiento de Hadoop?:

A) GFS.
B) DBFS.
C) HDFS.
D) FS.
ANSWER: C

2.- ¿Qué tipo de escalado es el más apropiado para Hadoop?:

A) Escalado vertical.
B) Escalado horizontal.
C) Escalado tangencial.
D) Hadoop no escala.
ANSWER: B

3.- ¿Cuál es la función del paso de "shuffle" en MapReduce?

A) Ordenar los datos de salida de los nodos "map" antes de enviarlos a los nodos "reduce".
B) Dividir los datos de entrada en trozos más pequeños para procesarlos de forma distribuida.
C) Realizar la operación de reducción en los datos de entrada.
D) Realizar una operación de mapeo en los datos de entrada.
ANSWER: A

4.- ¿Qué componente de Hadoop permite sincronizar el estado de los diferentes servicios distribuidos de Hadoop?

A) Apache Hive.
B) Apache ZooKeeper.
C) Apache HBase.
D) Apache Flume.
ANSWER: B

5.- ¿Qué comando HDFS se utiliza para copiar un archivo del sistema de archivos distribuido de Hadoop al sistema local?

A) hadoop fs -ls
B) hadoop fs -mkdir
C) hadoop fs -put
D) hadoop fs -get
ANSWER: D

6.- ¿Cómo consigue HDFS tener tolerancia a fallos?

A) Dividiendo los ficheros en bloques.
B) Almacenando los bloques en diferentes nodos.
C) Replicando los bloques en varios nodos.
D) Todas las anteriores.
ANSWER D

7.- En las primeras versiones de HADOOP cuál de estas afirmaciones NO es cierta

A) Restringía mucho el tipo de aplicaciones que los desarrolladores podían realizar.
B) La concurrencia en la ejecución de aplicaciones no estaba bien resuelta.
C) El procesamiento y el almacenamiento eran independientes.
D) Ninguna de las anteriores.
ANSWER C

8.- ¿En qué nodo se ejecuta el ApplicationMaster?

A) En el nodo maestro.
B) En el Datanode.
C) En un nodo worker.
D) En el ResourceManager.
ANSWER C

9.- ¿Cuáles son las fases de ejecución de un trabajo MapReduce?

A) map, suffle, reduce, order
B) map, order, suffle, reduce
C) map, suffle, order, reduce
D) map, reduce, suffle, order
ANSWER C

10.- ¿Cuál es el método que se utiliza para emitir pares clave-valor en el mapper de MRJob?

A) yield (key, value)
B) output.collect(key, value)
C) emit(key, value)
D) print(key, value)
ANSWER A

11.- Señala si son o no ciertas las siguientes afirmaciones:

I.- Las consultas de Hive se traducen a procesos MapReduce.
II.- Los datos en Hive se almacenan en una estructura relacional.

A) I cierta, II cierta
B) I cierta, II falsa
C) I falsa, II cierta
D) I falsa, II falsa
ANSWER B

12.- ¿Qué operador se utiliza para proyectar o seleccionar columnas de datos en Apache Pig?

A) SELECT
B) JOIN
C) COGROUP
D) FOREACH
ANSWER D

13.- Señala si son o no ciertas las siguientes afirmaciones sobre la siguiente consulta:

SELECT * FROM airports LIMIT 10
I.- La sintaxis es válida en Hive.
II.- La sintaxis es válida en Pig.

A) I cierta, II cierta
B) I cierta, II falsa
C) I falsa, II cierta
D) I falsa, II falsa
ANSWER B

14.- Señala si son o no ciertas las siguientes afirmaciones:

I.- Una transformación sobre un "dataframe" de Spark devuelve otro "dataframe".
II.- Una acción sobre un "dataframe" de Spark devuelve otro "dataframe".

A) I cierta, II cierta
B) I cierta, II falsa
C) I falsa, II cierta
D) I falsa, II falsa
ANSWER B

15.- ¿Cuál es el objetivo principal de Apache Ambari?

A) Añade capacidades de tiempo real en un clúster Hadoop.
B) Añade una capa de seguridad en un clúster Hadoop.
C) Simplifica la administración de clústeres Hadoop.
D) Ninguna de la anteriores.
ANSWER C

16.- El modelo tradicional basado en herramientas de ETL, Datawarehouses y herramientas de Business Intelligence y Data Mining es un buen modelo para realizar análisis de los datos para la toma de decisiones, pero tiene varios problemas asociados:

A) Sólo permite analizar datos estructurados, y cada vez hay un mayor número de fuentes de datos no estructuradas que se quieren analizar: logs de aplicaciones, transcripciones de conversaciones, imágenes, vídeos, etc.
B) Requiere mucha intervención de los equipos de tecnología o desarrollo, desde la construcción de los procesos ETL hasta la creación de los cuadros de mando o informes. Esto hace que desde que el negocio tiene una necesidad hasta que dispone de la herramienta para cubrir esa necesidad, el proceso puede durar demasiado tiempo.
C) A y B son ciertas.
D) A o B son falsas.
ANSWER C

17.- ¿Qué quiere decir la afirmación de que Hadoop es multitenancy?

I.- Que Hadoop permite que múltiples usuarios de diferente tipo utilicen la plataforma.
II.- Que Hadoop permite trabajar con muchos ficheros de cualquier tipo.

A) I cierta, II cierta
B) I cierta, II falsa
C) I falsa, II cierta
D) I falsa, II falsa
ANSWER A

18.- Señala si son o no ciertas las siguientes afirmaciones:

I.- La implementación de un Data Lake es más económica que la de un Datawarehouse.
II.- Un Datawarehouse sólo almacena datos estructurados, mientras que un Data Lake puede almacenar cualquier tipo de dato.

A) I cierta, II cierta
B) I cierta, II falsa
C) I falsa, II cierta
D) I falsa, II falsa
ANSWER A

19.- ¿Cuál de las siguientes afirmaciones NO es correcta en relación con los Data Lakes?

A) Hadoop es una buena plataforma para implementar un Data Lake.
B) Los Data Lakes intentan ser un repositorio de datos único para toda la empresa.
C) Frente a los Datawarehouses tradicionales, un Data Lake ofrece más funcionalidad.
D) Un Data Lake es más fácil de gestionar que un Datawarehouse.
ANSWER D

20.- ¿Cuál de las siguientes afirmaciones sobre EMR NO es correcta?

A) EMR permite arrancar clústers Hadoop rápidamente, por lo que es muy útil para hacer pruebas con Hadoop.
B) EMR permite configurar qué componentes del ecosistema Hadoop arrancar.
C) EMR puede adaptar el número de servidores a la carga real que esté soportando, por lo que sólo pagas por el uso real.
D) EMR, que son las siglas de Elastic MapReduce, sólo permite MapReduce como framework para procesar datos.
ANSWER D

Modelo 2.

1.- ¿Cómo se llama la disciplina que tradicionalmente ha creado modelos predictivos sobre los datos del Datawarehouse?

A) Data Mining
B) Data Exploration
C) Data Discovery
D) ETL
ANSWER A

2.- ¿Cómo se consigue escalar o ganar mayor capacidad en una plataforma Hadoop?

A) Cambiando los servidores por otros con CPUs más potentes
B) Añadiendo más nodos master al clúster
C) Realizando optimizaciones en los sistemas operativos
D) Añadiendo más nodos worker al clúster
ANSWER D

3.- ¿Cuál de las siguientes afirmaciones es más apropiada para Hadoop?

A) Hadoop se despliega en infraestructura propia, no en entornos cloud
B) Cambió el paradigma tradicional, acercando el procesamiento a donde se almacenan los datos
C) Es la mejor tecnología para cualquier caso de uso Big Data
D) El nivel de seguridad es muy alto, al nivel de otras tecnologías de gestión de datos tradicionales (como las bases de datos relacionales)
ANSWER B

4.- ¿Cómo se llaman las herramientas que preparan los datos para el Datawarehouse?

A) El Datawarehouse no requiere una preparación previa de los datos
B) Herramientas de ETL
C) Herramientas de Data Governance
D) Herramientas de Data Discovery
ANSWER B

5.- Si quiero arrancar un Hadoop en la nube como servicio y quiero usar Ambari para gestionar el clúster, ¿qué solución debería utilizar?

A) Cloudera
B) HDInsight
C) Ninguna lleva Ambari
D) EMR
ANSWER B

6.- ¿A qué se refiere el concepto "industrialización" en relación con las actividades de ingeniería de datos?

A) A llevar todas las plataformas a la nube
B) A estandarizar y automatizar todo lo posible las tareas para ganar eficiencia
C) A aplicar tecnologías Big Data en el sector industrial
D) A aplicar mecanismos de control y validación de la calidad
ANSWER B

7.- ¿Cómo se llama el principal fichero de configuración para el servicio HDFS?

A) core-config.xml
B) hadoop-commons.xml
C) hdfs-config.xml
D) hdfs-site.xml
ANSWER D

8.- ¿Qué tipo de datos gestiona un Datawarehouse?

A) Datos no estructurados
B) Datos estructurados
C) Datos semi-estructurados
D) Cualquier tipo de datos
ANSWER B

9.- ¿Cómo se llama el principal fichero de configuración para el servicio Hive?

A) hive-site.xml
B) hive-config.xml
C) hadoop-commons.xml
D) core-config.xml
ANSWER A

10.- ¿Qué componente del ecosistema Hadoop permite ver los ficheros HDFS como si fueran tablas de una base de datos relacional?

A) YARN
B) Oozie
C) Hive
D) Ambari
ANSWER C

11.- ¿Cuál es la sentencia de Hive con la que se borra una nueva tabla?

A) DELETE TABLE
B) DROP TABLE
C) DEL TABLE
D) FORMAT TABLE
ANSWER B

12.- ¿Cómo se llama el componente que ofrece, desde una web, acceso a los ficheros de HDFS y poder lanzar consultas Hive?

A) Hue
B) Hive
C) Impala
D) Spark
ANSWER A

13.- ¿Cuál es la sentencia de Hive con la que modificamos los registros de una tabla?

A) MODIFY ROWS
B) UPDATE
C) LOAD DATA
D) SELECT
ANSWER B

14.- ¿Cuál de las siguientes afirmaciones sobre Impala es correcta?

A) Permite acceder a los datos de HDFS como si fuera una tabla
B) Ninguna de las anteriores es correcta
C) Permite importar datos de otras bases de datos relacionales
D) Permite administrar un clúster Hadoop
ANSWER A

15.- Para ingestar datos que están en bases de datos relacionales, ¿qué componente de Hadoop se utiliza?

A) Pig
B) Sqoop
C) Flume
D) YARN
ANSWER B

16.- ¿Qué componente del ecosistema Hadoop fue el primero en aparecer para reducir la complejidad de los procesos MapReduce que se desarrollaban hasta entonces?

A) Spark
B) Pig
C) HDFS
D) YARN
ANSWER B

17.- Para automatizar la ejecución de trabajos que se debe realizar en Hadoop, por ejemplo, para validar los datos ingestados, ¿qué componente se debe utilizar?

A) YARN
B) Pig
C) Zookeeper
D) Oozie
ANSWER D

18.- ¿Qué componente del ecosistema Hadoop permite utilizar sintaxis SQL para manejar datos que están almacenado en HBase?

A) Oozie
B) Phoenix
C) Storm
D) Pig
ANSWER B

19.- Para ingestar datos que se generan en tiempo real, ¿qué componente de Hadoop se utiliza?

A) Storm
B) Spark
C) Flink
D) Flume
ANSWER D

20.- ¿Cómo se llama el principal fichero de configuración para el servicio YARN?

A) yarn-config.xml
B) yarn-site.xml
C) hadoop-commons.xml
D) core-config.xml
ANSWER B