Exámenes

Test de Evaluación de CBD

Volver a la Lista de Exámenes

Parte 1 de 1 - Preguntas 9.5 / 10.0 Puntos

Preguntas 1 de 20 0.5
0.5 Puntos

¿Cuál de los siguientes proyectos/herramientas no forma parte del ecosistema de Hadoop?

- ✔ ○ A. Pig
- ✔ ○ B. Ansible
- ✔ ○ C. Hive
- ✔ ○ D. Cassandra

Respuesta correcta: B

Comentarios: Pig, Hive y Cassandra son proyectos del ecosistema de Hadoop. Sin embargo, Ansible es una herramienta de automatización de tipo DevOps para la configuración desatendida de infraestructuras.

Preguntas 2 de 20 0.5
0.5 Puntos

Se despliega un cluster Hadoop con Amazon EMR con un nodo Máster, 7 nodos Core y 3 nodos Task. El nodo Máster está basado en spot instances (instancias puntuales) de manera que, en un momento dado, la instancia EC2 asociada al nodo Máster se termina. ¿Qué ocurre con el cluster?

- ✔ ○ A. Amazon EMR despliega un nuevo nodo para reemplazar al nodo Máster, cogiendo los datos de copia de seguridad de HDFS
- ✔ ○ B. El cluster queda inservible.
- ✔ ○ C. No es posible desplegar un nodo Máster basado en spot instances.
- ✔ ○ D. Uno de los nodos Core promociona automáticamente a Máster y asume el control de las tareas pendientes.

Respuesta correcta: B

Comentarios: El nodo Máster ejecuta el servicio NameNode de Hadoop que lleva una asignación de qué bloques están almacenados en qué nodos. Además, únicamente se permite desplegar un nodo Máster y éste

puede estar basado en spot instances, de manera que es posible pujar por el precio al que se quiere desplegar dicha instancia. Una vez desplegada la instancia, si el precio sube por encima del precio de puja, la instancia será terminada y, por tanto, al no haber nodo Máster, el cluster quedará inaccesible. Amazon EMR no implementa ningún esquema de tolerancia a fallos en caso de fallo del nodo máster (en caso de fallo de un nodo Core despliega automáticamente otro nodo).

Preguntas 3 de 20 0.5

0.5 Puntos

El modelo de programación MapReduce se utiliza principalmente para ...

- ✔ ○ A. Ejecutar programas computacionalmente complejos sobre conjuntos de datos de almacenamiento centralizado.
- ✔ ○ B. Ejecutar programas paralelos con librerías de paso de mensajes como OpenMPI.
- ✔ ○ C. Ejecutar trabajos que procesan datos almacenados en memoria principal.
- ✔ ○ D. Ejecutar tareas que procesan datos sobre conjuntos de datos de gran dimensión almacenados en memoria secundaria (disco).

Respuesta correcta: D

Comentarios: MapReduce es un modelo de programación que permite la ejecución de trabajos sobre conjuntos de datos de gran dimensión almacenados en un sistema de archivos distribuido (HDFS).

Preguntas 4 de 20 0.5

0.5 Puntos

¿Para qué sirve el servicio Amazon EMR?

- ✔ ○ A. Para ejecutar programas MapReduce sin necesidad de desplegar un cluster.
- ✔ ○ B. Para facilitar el despliegue de clusters de cómputo para procesamiento de datos en AWS.
- ✔ ○ C. Para facilitar la creación de programas MapReduce mediante un lenguaje de alto nivel.
- ✔ ○ D. Para analizar las prestaciones de programas MapReduce y desplegar clusters Hadoop on-premises customizados para dichos programas.

Respuesta correcta: B

Comentarios: Amazon EMR facilita el despliegue de clusters de cómputo para procesamiento de datos en AWS, posibilitando el uso de herramientas como Hadoop, Spark, Pig, Hive, Hue, etc. No permite ejecutar programas sin desplegar un cluster, no ofrece un lenguaje de alto nivel adicional al ofrecido por las herramientas que instala y tampoco se encarga de analizar las prestaciones de programas MapReduce.

Preguntas 5 de 20 0.5

0.5 Puntos

¿Qué tipos de nodos pueden existir en un cluster desplegado con Amazon EMR? Elige todas las opciones correctas.

- [ ] A. DataNode
- [x] B. Task
- [x] C. Core
- [x] D. Master
- [ ] E. Executor
- [ ] F. NameNode

Respuesta correcta: B, C, D

Comentarios: Un cluster desplegado con Amazon EMR soporta tres tipos diferentes de nodos. i) Un nodo Máster encargado de asignar tareas a nodos Core y nodos Task así como monitorizar su estado; ii) nodos Core que ejecutan tareas y almacenan datos en HDFS y iii) nodos Task que únicamente ejecutan tareas (no almacenan datos en HDFS).

Preguntas 6 de 20 0.5
0.5 Puntos

¿Para qué tipo de operaciones de E/S está particularmente optimizado HDFS (Hadoop Distributed File System)?

- [x] A. Escrituras aleatorias largas
- [x] B. Lecturas secuenciales largas
- [x] C. Escrituras secuenciales cortas
- [x] D. Lecturas aleatorias cortas

Respuesta correcta: B

Comentarios: El sistema de archivos distribuido de Hadoop (HDFS) está especialmente optimizado para lecturas secuenciales largas. Lecturas aleatorias y, particularmente, escrituras aleatorias son más costosas en tiempo de acceso a los datos.

Preguntas 7 de 20 0.5
0.5 Puntos

¿Para qué sirve Hadoop Streaming?

- [x] A. Para crear programas MapReduce en cualquier lenguaje de programación.
- [x] B. Para conseguir lecturas más eficientes orientadas a flujo.
- [x] C. Para ejecutar workflows de tareas (streamline) de programas MapReduce.
- [x] D. Para realizar escrituras orientadas a flujo eficiente desde un programa MapReduce.

Respuesta correcta: A

Comentarios: Hadoop Streaming permite la creación y ejecución de trabajos MapReduce implementados con cualquier lenguaje de programación, ya que el único requisito es que los programas lean y escriban de la salida estándar.

Preguntas 8 de 20 0.5
0.5 Puntos

¿Cuál es el mecanismo de detección de transición entre claves para la fase de Reduce de un código MapReduce implementado en Python para poder utilizar Hadoop Streaming?

- ✔ ○ A. El script que implementa el Reduce asume que los pares clave/valor vienen ordenados lexicográficamente por el valor, a través de la entrada estándar.
- ✔ ○ B. La función Reduce recibe como parámetro tantos arrays como claves tenga el fichero de entrada.
- ✔ ○ C. El script que implementa el Reduce asume que los pares clave/valor vienen ordenados lexicográficamente por la clave, a través de la entrada estándar.
- ✔ ○ D. No entiendo la pregunta.

Respuesta correcta: C

Comentarios: Un programa MapReduce implementado en Python para poder utilizar Hadoop Streaming en un cluster Hadoop involucra dos ficheros, uno para la fase de Map y otro para la fase de Reduce. El script que implementa la fase de Reduce debe asumir que los pares clave/valor emitidos durante la fase de Map vendrán ordenados por clave y serán accesibles a través de la entrada estándar. Esto permitirá al código detectar la transición entre claves puesto que los pares estarán ordenados y se podrá detectar cuando se comienzan a obtener pares correspondientes a una clave diferente.

Preguntas 9 de 20 0.5
0.5 Puntos

Es posible crear un programa MapReduce que se ejecute sobre un cluster Hadoop sin necesidad de escribir código Java. ¿Verdadero o Falso?

✔ ○ Verdadero
○ Falso

Respuesta correcta: Verdadero

Comentarios: Hadoop Streaming permite la creación de programas MapReduce que leen y escriben en la salida estándar, por lo que no es necesario que estén programados en Java. Por ejemplo, pueden estar escritos en otro lenguaje de programación (como Python, Ruby o C) y la única restricción es que reciban los datos por la entrada estándar y emitan los datos por la salida estándar.

Preguntas 10 de 20 0.5

0.5 Puntos

¿Cuál es la característica principal del sistema de archivos de Hadoop (HDFS)?

- ✔ ○ A. Centraliza el almacenamiento de datos en un solo nodo por cuestiones de eficiencia.
- ✔ ○ B. Almacena bloques de 512 MB con 7 replicas a lo largo de los nodos de un cluster Hadoop.
- ✔ ○ C. Almacena bloques de 64 MB con 3 replicas en diferentes nodos de un cluster Hadoop.
- ✔ ○ D. Puede federar automáticamente datos de múltiples clusters Hadoop.

Respuesta correcta: C

Comentarios: HDFS (Hadoop Distributed File System) es un sistema de archivos distribuido que, por defecto, particiona los ficheros en bloques de 64 MB y almacena 3 copias de los mismos en diferentes nodos de un cluster Hadoop.

Preguntas 11 de 20 0.5

0.5 Puntos

¿Qué mecanismo utiliza Apache Spark para resolver el problema de caída/fallo de un nodo del cluster durante la ejecución de un programa?

- ✔ ○ A. Se realiza persistencia periódica en HDFS para tener copia de seguridad
- ✔ ○ B. Ninguno. Se detiene la ejecución del programa con un fallo.
- ✔ ○ C. Se ejecutan nuevamente las tareas dependientes previas necesarias para poder obtener la información que estaba en el nodo que ha fallado.
- ✔ ○ D. Se realiza persistencia periódica en la memoria RAM de otros nodos para evitar fallos en un nodo.

Respuesta correcta: C

Comentarios: Apache Spark utiliza el concepto de RDD (Resilient Distributed Dataset) para disponer de datasets particionados e inmutables sobre los que se conoce la secuencia de operaciones ejecutada para llegar a él (lineage). En caso de fallo de un nodo se deben ejecutar nuevamente las operaciones previas para volver a regenerar dicha partición en un nuevo nodo y proseguir la ejecución del programa.

Preguntas 12 de 20 0.5

0.5 Puntos

En una ejecución de un programa MapReduce, la fase de Reduce requiere que previamente se haya realizado una fase de ordenación por clave. ¿Verdadero o Falso?

✔ ○ Verdadero

○ Falso

Respuesta correcta: Verdadero

Comentarios: En MapReduce, tras la fase de Map se realiza la fase Shuffle & Sort encargada de ordenar por la clave los datos obtenidos como resultado de la fase Map. Los resultados ordenados pasan a continuación a la fase de Reduce.

Preguntas 13 de 20 0.5
0.5 Puntos

¿Cuales son las fases principales de la ejecución de un programa MapReduce sobre un cluster Hadoop?

- ✔ A. Map, Shuffle & Sort y Reduce
- ✔ B. Spread, Map, Shuffle & Sort, Reduce, Store
- ✔ C. Map, Shuffle & Sort, Reduce, Store
- ✔ D. Map y Reduce.

Respuesta correcta: A

Comentarios: La ejecución de un programa MapReduce sobre un cluster con Apache Hadoop involucra varias fases pero las principales son Map, que emite pares clave valor, Shuffle & Sort, donde se agrupan los pares por clave y Reduce, donde se aplica una operación a todos los valores de una misma clave, para todas las claves.

Preguntas 14 de 20 0.0
0.5 Puntos

¿Es posible redimensionar (en número de nodos) un cluster Hadoop?

- ✔ A. Sí
- ✔ B. No
- ✖ C. Depende del espacio libre en el sistema de archivos HDFS
- ✔ D. Depende del tiempo transcurrido desde que se desplegó.

Respuesta correcta: A

Comentarios: Sí es posible aumentar y reducir el número de nodos de un cluster Hadoop. Por ejemplo, Amazon EMR permite el despliegue de clusters Hadoop y, posteriormente, ampliar o reducir el número de nodos asignado al cluster de manera que se puede ajustar el tamaño del mismo a las necesidades de cómputo en cada momento.

Preguntas 15 de 20 0.5
0.5 Puntos

¿Qué operación es más lógica utilizar en un código a ejecutar en PySpark para poder obtener un subconjunto de líneas en las que aparezca una determinada palabra a partir de un dataset ya pre-cargado con la operación textFile?

- ✔ ◯ A. take
- ✔ ◯ B. reduceByKey
- ✔ ◯ C. map
- ✔ ◯ D. filter
- ✔ ◯ E. split

Respuesta correcta: D

Comentarios: La operación más lógica es filter, que permite crear un nuevo RDD a partir de un RDD ya creado de manera que contenga únicamente un subconjunto de líneas del RDD original (aquellas que contengan dicha palabra).

Preguntas 16 de 20 0.5
0.5 Puntos

El paradigma de computación HTC (High Throughput Computing) ...

- ✔ ◯ A. Se basa en aplicaciones paralelas que deben aprovechar al máximo las capacidades de los múltiples procesadores del sistema paralelo.
- ✔ ◯ B. Se basa en el uso de trabajos que intercambian mensajes entre ellos para facilitar el progreso de la ejecución.
- ✔ ◯ C. Se basa en múltiples trabajos independientes entre sí que deben ser ejecutados el mayor número de ellos en el menor tiempo posible.
- ✔ ◯ D. Requiere el uso de una arquitectura de tipo cluster de PCs para su ejecución.

Respuesta correcta: C

Comentarios: El paradigma de computación HTC (High Throughput Computing) se basa en múltiples trabajos independientes entre sí de manera que no requieren realizar ninguna comunicación entre ellos. Esto permite ejecutarlos de forma simultánea sobre infraestructuras distribuidas (como un Grid) o locales (como un cluster) de manera que puedan ser ejecutados lo más rápido posible el máximo número de ellos.

Preguntas 17 de 20 0.5
0.5 Puntos

¿Cuál de las siguientes opciones consideras que es la principal ventaja de Apache Spark?

- ✔ ◯ A. Permite la integración con fuentes de datos de múltiples proveedores Cloud para poder acceder a múltiples bases de datos.

- ✔ ○ B. Permite acelerar el proceso de escritura en el sistema de archivos distribuido HDFS.
- ✔ ○ C. Permite la ejecución de programas MapReduce creados en múltiples lenguajes de programación.
- ✔ ○ D. Permite almacenar resultados en memoria para reducir el número de accesos a HDFS.

Respuesta correcta: D

Comentarios: Apache Spark incluye dos ventajas fundamentales frente a Apache Hadoop. En primer lugar, permite almacenar resultados parciales en memoria para minimizar el número de accesos a HDFS y agilizar la computación. En segundo lugar, permite optimizar el grafo de dependencias entre tareas para optimizar la ejecución de las tareas.

Preguntas 18 de 20 0.5
0.5 Puntos

Elige cual de las siguientes opciones NO es una de las características de los datos que pretenden ser procesados mediante técnicas de Big Data.

- ✔ ○ A. Velocidad a la que se producen los datos.
- ✔ ○ B. Verbosidad o sobre-exceso de datos.
- ✔ ○ C. Volumen o tamaño del conjunto de datos.
- ✔ ○ D. Variedad de formatos y fuentes de las que se obtienen los datos.

Respuesta correcta: B

Comentarios: Los datos que pretenden ser analizados mediante técnicas de Big Data suelen cumplir todas o alguna de las siguientes características: i) Volumen, o tamaño del conjunto de datos; ii) Velocidad, a la que producen los datos y iii) Variedad, de fuentes y de formatos en los que se encuentran u obtienen dichos datos.

Preguntas 19 de 20 0.5
0.5 Puntos

¿Para qué sirve principalmente Apache Hue?

- ✔ ○ A. Es una interfaz web que permite simplificar la interacción con un cluster Hadoop para el usuario.
- ✔ ○ B. Es una interfaz web que permite obtener métricas de uso del cluster Hadoop, siempre que haya sido desplegado con Amazon EMR.
- ✔ ○ C. Es un framework que incluye un planificador de tareas optimizado para substituir a YARN, el scheduler por defecto de Apache Hadoop.

- ✔ ○ D. Es una interfaz REST que permite interconectar Apache Hadoop con otros servicios de AWS de analítica de datos.

Respuesta correcta: A

Comentarios: Apache HUE (Hadoop User Experience) es una aplicación web de código abierto que permite facilitar el uso de un cluster Hadoop para usuarios finales. Permite servir como punto de entrada a diferentes herramientas del ecosistema de Hadoop (como por ejemplo Apache Pig y Apache Hive). También facilita la importación de datos al sistema de archivos distribuido HDFS.

Preguntas 20 de 20 0.5

0.5 Puntos

La principal ventaja de disponer de un cluster virtual en la nube, desplegado sobre un proveedor de Cloud público es ...

- ✔ ○ A. Eliminar completamente los gastos de operación del cluster.
- ✔ ○ B. Reducir la complejidad de administración del cluster una vez ha sido desplegado.
- ✔ ○ C. Evitar el desembolso inicial de adquisición del cluster y pagar únicamente por los recursos utilizados.
- ✔ ○ D. El aumento automático de las prestaciones de las aplicaciones.

Respuesta correcta: C

Comentarios: La principal ventaja es evitar el coste de amortización derivado del desembolso inicial de adquisición de un cluster y pagar únicamente por los recursos utilizados cuando sean utilizados. En un cluster virtual siguen habiendo gastos de operación derivados del consumo de recursos de cómputo del proveedor de Cloud público. Tampoco se reduce la complejidad de administración del cluster puesto que una vez desplegada, sea virtual o físico, hay que actualizar los parches de seguridad, gestionar los usuarios, etc. Finalmente, el uso de un cluster virtual no determina por si solo un aumento de las prestaciones de las aplicaciones, ya que eso dependerá de las características de los recursos virtuales utilizados, de la capacidad de la red entre ellos, etc.