

Herramientas de Big Data

Preguntas Tipo Examen

1. Según su opinión qué problemas de su vida laboral podrían ser resueltos usando técnicas y/o herramientas de Big Data
2. Indicar el significado de cada una de las 3 Vs (principales) y explicar cómo es que influyen cada una a la hora de plantear que los problemas de Big Data están en una nueva categoría de problemas de manejo de información.
3. ¿Por qué el acceso a los datos es la principal necesidad para una empresa?
4. Indique cuál es la función de cada una de las 3 capas de la Arquitectura Lambda e indique ejemplos de herramientas que se podrían usar en cada una.
5. Con respecto al procesamiento de grandes volúmenes de información ¿Qué beneficios trae que actualmente el hardware sea barato y hasta se pueda adquirir vía cloud?
6. ¿Cuáles son las principales características de HDFS?
7. En HDFS cuál es la función de estos componentes:
 - a. Name Node
 - b. Data Node
8. ¿Cuáles son las limitantes de HDFS?
9. Explicar cuál es la finalidad de cada una de las siguientes etapas en Hadoop
 - a. Shuffle
 - b. Reduce
10. ¿Qué tipo de ganancia se puede obtener al implementar un Combiner en Hadoop?
11. ¿Qué problemas se intentan resolver en Hadoop con la creación de YARN?
12. ¿Cuáles son las principales características generales que definen a las Bases NoSQL?
13. ¿Cuáles son los 3 componentes del teorema CAP y que indica el teorema de la relación entre ellos para sistemas distribuidos?
14. ¿Qué es lo que se indica al decir que una base cumple con consistencia eventual?
15. ¿Las bases de datos Columnares por su manera de guardar son muy eficientes para qué tipo de consultas?
16. ¿Cuál es la diferencia entre una base de datos key value y una de tipo documento en cuanto al tratamiento del "valor" (pensando que una documento puede considerarse que el documento es el valor)?
17. ¿Por qué HBASE logra tener consistencia de escritura a pesar de ser distribuido?
18. ¿Indicar 3 de los casos de uso comunes para la utilización de REDIS?
19. ¿Cuál es la diferencia entre una tabla declarada externa y una que no fue declarada así en Hive?
20. ¿Donde guarda Hive el contenido de las tablas y donde la metadata (que tablas, que columnas, donde se guardó la data, etc)?
21. ¿Qué beneficios trae conocer el esquema de la data a la hora de guardarlas y realizar consultas sobre la misma?
22. ¿Cuáles son las características principales de los RDDs?

23. Cuales son los dos tipos de operaciones que se pueden realizar sobre los RDDs y que provocan cuando se los invoca.
24. Indicar las ventajas y desventajas provee el poder cachear (persistir) un RDD.
25. Dado un set de datos compuestos por ---...--- Resolver (sin codificar) la query ---...--- via map reduce indicando:
- La clave y el valor de salida del mapper (indicando si hubo filtrado o transformación).
 - La operación a realizar en el proceso de reduce.
 - La clave y valor de salida para el proceso de reduce.
26. Dado un set de datos compuesto por ---...--- Resolver vía pig la siguiente query: ---...--- Utilizando los siguientes operadores de pig (de necesitarlos) LOAD, FOREACH, FILTER, GROUP, ORDER, STORE.
27. Dado un set de datos s compuesto por ---...--- Resolver usando spark