Herramientas de Big Data
Preguntas Tipo Examen

1. Según su opinión qué problemas de su vida laboral podrían ser resuelto 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. ¿Que beneficios trae que actualmente el hardware sea barato y hasta se pueda adquirir vía cloud?
6. ¿Porque es importante implementar herramientas de despliegue automático para el mantenimiento de herramientas clusterizados?
7. ¿Cuales son las principales características de HDFS?
8. En HDFS cual es la función de estos componentes:
    a. Name Node
    b. Data Node
9. ¿Cuales son las limitantes de HDFS?
10. Explicar cuál es la finalidad de cada una de las siguientes etapas en Hadoop
    a. Shuffle
    b. Reduce
11. ¿Qué tipo de ganancia se puede obtener al implementar un Combiner en Hadoop?
12. ¿Cuales son los principales puntos a tener en cuenta para trabajar eficientemente con Hadoop?
13. ¿Qué problemas se intentan resolver en Hadoop con la creación de YARN?
14. ¿Cuales son las principales carácteristicas generales de las Bases NoSQL?
15. ¿Cuales son los 3 componentes del teorema CAP y que indica el teorema de la relación entre ellos para sistemas distribuidos?
16. ¿Qué es lo que se indica al decir que una base cumple con consistencia eventual?
17. ¿En bases NoSQL aumentar el factor de replicación como afecta a la latencia y la consistencia eventual?
18. ¿Las bases de datos Columnares por su manera de guardar son muy eficientes para qué tipo de consultas?
19. ¿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)?
20. ¿Por qué HBASE logra tener consistencia de escritura a pesar de ser distribuido?
21. Explicar la estrategia de Cassandra realiza la escritura y replicación de un dato.
22. ¿Cual es la restricción que presenta Cassandra a la hora de realizar filtros en sus consultas, y qué desafío se presenta a partir de esta restricción?
23. ¿Indicar alguno de los casos de uso más comunes para la utilización de REDIS?

24. El tratamiento del valor en REDIS es un poco más avanzado que en una base de datos key value “pura” ¿Por qué?
25. ¿Qué característica aporta ElasticSearch al mundo NoSQL
26. ¿Qué problema resuelve SQOOP?
27. ¿Qué características debe cumplir una base de datos para que sqoop pueda importala completa?
28. ¿Cuáles son los dos modos de importar incrementalmente vía SQOOP a HDFS?
29. En Flume cual es la responsabilidad de estos componentes:
    a. Source
    b. Channel
    c. Sink
30. ¿Cuál es la diferencia entre una tabla declarada external y una que no fue declarada así en Hive?
31. ¿Donde guarda Hive el contenido de las tablas y donde la información de las mismas (la metadata)?
32. Comparar Hive vs. Pig como herramientas de procesamiento mencionando las ventajas de cada una y por cuál se inclinaría.
33. ¿Qué beneficios trae conocer el esquema de la data a la hora de guardarlas y realizar consultas sobre la misma?
34. ¿Cuales son las características principales de los RDDs?
35. Indicar los componentes internos que definen a un RDD.
36. Para spark existen dos tipos de grupos de operaciones, indicar cuales son, cual es lazy y cual es eager, y que significan esos conceptosCuales son los dos tipos de operaciones que se pueden realizar sobre los RDDs y que provocan cuando se los invoca.
37. Indicar las ventajas y desventajas provee el poder cachear (persistir) un RDD.
38. Cuales son las estructuras de dato que SparkSQL y que ventajas tiene sobre un “simple” RDD.
39. 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.
40. 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.