Cuestionario de PySpark

1. **¿PySpark soporta leguajes de programación cómo?**
   - Python
   - JavaScript
   - Java
   - Go

2. **¿Cuántos componentes tiene la arquitectura de PySpark?**
   - 3
   - 5
   - 2

3. **¿Qué hace el Cluster Manager?**
   - Administrar la asignación de recursos.
   - Administrar las versiones de cada componente.
   - Administrar la división del programa.
   - Administrar la ejecución del programa.

4. **¿Qué hacen los Executors?**
   - Es el proceso principal, controla toda la aplicación y la ejecuta.
   - Hace referencia a las máquinas que dependen del backend.
   - Adquiere recursos físicos y poder ejecutar los executors.

5. **¿Qué hace el Driver?**
   - Permite que se extraiga la información de los sistemas locales y se desarrolle en sistemas distribuidos.
   - Despliega en una de las capas de control que se instaura sobre las aplicaciones o comandos que corren sobre Hadoop.
   - Es el proceso en el que se realiza la carga de trabajo, de manera que obtienen sus tareas desde el driver para cargar, transformar y almacenar los datos.

6. **¿Qué es PySpark?**
   - Una librería de Python para el procesamiento de datos en grandes volúmenes.
   - Un sistema de gestión de bases de datos.
   - Un lenguaje de programación para el desarrollo web.

7. **¿PySpark no es solo para el procesamiento de grandes volúmenes de datos?**
   - Verdadero.
   - Falso.

8. ¿PySpark es compatible con qué versión de Python?

- Python 2.7
- Python 3.x
- Todas las anteriores.

9. ¿PySpark es compatible con Hadoop?

- Verdadero
- Falso

10. ¿PySpark puede ser utilizado para procesar datos en tiempo real?

- Verdadero
- Falso

11. Seleccione una desventaja de PySpark

- Generalmente se considera difícil.
- Permite el uso de RDD para tolerancia a fallas.
- Es un lenguaje interpretado

12. PySpark es una herramienta que admite Python en Spark

- Verdadero
- Falso

13. Un Dataframe en Pyspark se construye sobre la base del RDD.

- Verdadero
- Falso

14. Cuál es la definición de Spark Core

- Una colección de elementos que es tolerante a fallos y que es capaz de operar en paralelo.
- Es la base para todo el procesamiento de datos en paralelo y maneja la programación, optimización, RDD y abstracción de datos.
- Es un componente de Spark que admite el procesamiento escalable y tolerante a errores de transmisión de datos.

15. Seleccione una característica de los RDD (Resilient Distributed Dataset)

- Interpretado
- Productividad
- Inmutables

16. **Los DataFrames permiten a PySpark consultar los datos de dos maneras diferentes: SQL, por ejemplo (SELECT * from table) y método de expresión, por ejemplo (df.select()).**
    - Verdadero.
    - Falso.

17. **Algunos operadores de DataFrame en PySpark correspondientes a las acciones son:**
    - count()
    - filter()
    - columns
    - groupby()

18. **¿Qué es PySpark MLlib?**
    - Es una extensión de la API central de Spark que permite el procesamiento de flujos de datos en vivo escalable, de alto rendimiento y tolerante a fallas.
    - Es una librería integrada para Machine Learning escalable.
    - Es una biblioteca Spark para datos estructurados.
    - es un paquete para Apache Spark que proporciona gráficos basados en DataFrames.

19. **¿Un ejemplo de PySpark Streaming es recomentar el correcto programa de TV o película utilizando pyspark.streaming.StreamingContext()?**
    - Verdadero.
    - Falso.

20. **Algunas aplicaciones de PySpark GraphFrames son las siguientes:**
    - Motores de búsqueda.
    - Análisis de tendencias.
    - Redes sociales.
    - Detección de fraudes.
    - Inteligencia Artificial.
    - Análisis bursátil en tiempo real.