

**1. ¿Cuál de las siguientes plataformas de Microsoft es una solución de análisis de big data en la nube?**

- a) Azure SQL Database
- b) Azure Synapse Analytics
- c) Azure Data Factory
- d) Azure Active Directory

**Respuesta:** Azure Synapse Analytics

Azure Synapse Analytics integra almacenamiento de datos y análisis de big data en un único servicio gestionado, facilitando consultas a gran escala y procesamiento distribuido en la nube.

**2. En el contexto de Azure Data Factory, ¿qué actividad se utiliza para transformar y limpiar datos en un flujo de trabajo?**

- a) HDInsight Spark
- b) Azure Databricks
- c) Data Flow
- d) Azure Stream Analytics

**Respuesta:** Data Flow

Data Flow es la actividad nativa de Azure Data Factory diseñada para aplicar transformaciones visuales y ETL sin escribir código, ideal para limpieza y modelado de datos en pipelines.

**3. ¿Qué característica clave de Apache Spark permite procesar datos en memoria para un rendimiento más rápido?**

- a) Apache Hadoop
- b) Apache Flink
- c) Spark Streaming
- d) Resilient Distributed Dataset (RDD)

**Respuesta:** Resilient Distributed Dataset (RDD)

Los RDDs son la abstracción fundamental de Spark que almacena particiones de datos en memoria y permite operaciones de transformación perezosa, reduciendo lecturas/escrituras en disco y acelerando los cálculos.

**4. En Pandas, ¿qué operación se utiliza para eliminar filas duplicadas de un DataFrame?**

- a) df.groupby()
- b) df.drop\_duplicates()
- c) df.fillna()
- d) df.pivot\_table()

**Respuesta:** df.drop\_duplicates()

El método drop\_duplicates() remueve automáticamente filas repetidas basándose en todas las columnas (o en un subconjunto) y devuelve un DataFrame sin duplicados.

**5. ¿Qué lenguaje de programación se utiliza comúnmente en Azure Databricks para el procesamiento de datos y análisis?**

- a) R
- b) Java
- c) Scala
- d) C#

**Respuesta:** Scala

Azure Databricks se construye sobre Apache Spark, cuyo API nativo está en Scala; por ello, Scala suele ofrecer el mejor rendimiento y la compatibilidad más directa para flujos de trabajo en Databricks.