Tema 1. Introducción a las tecnologías big data

1. En la sociedad actual, la mayoría de los datos que se generan a diario son… A. Datos no estructurados generados por las personas. -> **correcto** B. Datos estructurados generados por máquinas. C. Datos estructurados generados por las personas.

2. ¿Qué retos presentan los datos generados por personas en una red social? A. Son datos no estructurados (imágenes, vídeos), más difíciles de procesar. B. Son datos masivos. C. Las dos respuestas anteriores son correctas. -> **Correcto**

3. El término commodity hardware se refiere a… A. Máquinas remotas que se alquilan a un proveedor de cloud como Amazon. B. Máquinas muy potentes que suelen adquirir las grandes empresas. C. Máquinas de potencia y coste normales, conectadas entre sí para formar un clúster más potente. -> **Correcto**

4. Un proyecto se denomina big data cuando… A. Solo se puede resolver gracias a las tecnologías big data. B. La forma más eficaz y directa de abordarlo implica tecnologías big data.-> **Correcto** C. El problema que resuelve contiene simultáneamente las tres «v».

5. Las tres «v» del big data se refieren a: A. Volumen, velocidad y variedad. -> **Correcto** B. Voracidad, volumen y velocidad. C. Ninguna de las respuestas anteriores es correcta.

6. Lo mejor, si necesitamos más potencia de cómputo en un clúster big data, es... A. Reemplazar algunas máquinas del clúster por otras más potentes. B. Aumentar el ancho de banda de la red. C. Añadir más máquinas al clúster y aprovechar todas las que ya había. -> **Correcto**

7. El sistema de ficheros precursor de HDFS fue... A. GFS. -> **Correcto** B. Apache Hadoop. C. Apache MapReduce.

8. Una distribución de Hadoop es... A. Un software con licencia comercial para clústeres, difundido por Microsoft. B. Un conjunto de aplicaciones del ecosistema Hadoop, con versiones interoperables entre sí y listas para usarse. -> **correcto** C. Ninguna de las opciones anteriores es correcta.

9. ¿Qué compañías fueron precursoras de HDFS y MapReduce? A. Google y Microsoft, respectivamente. B. Google, en los dos casos. -> **correcto** C. Google y Apache, respectivamente.

10. Definimos big data como... A. Todos aquellos algoritmos que se pueden ejecutar sobre un clúster de ordenadores. B. Las tecnologías distribuidas de Internet que posibilitan una sociedad interconectada por las redes sociales. C. Las tecnologías que permiten almacenar, mover, procesar y analizar cantidades inmensas de datos heterogéneos. -> **correcto**

Videoclase 1. Sociedad de la información

- ¿Cuál es uno de los desafíos principales en la era de la sociedad interconectada?
- La capacidad de asimilar y almacenar el flujo constante de datos en tiempo real.
- ¿Qué permite la utilización de tecnologías de Big Data en las empresas?
- Almacenar y procesar grandes volúmenes de datos para tomar decisiones informadas.
- ¿Cuál es un beneficio significativo de la interacción humano-máquina?
- Mejora de la accesibilidad para personas con discapacidades.
- ¿Cuál es un impacto positivo del uso de Big Data en la atención médica?
- Personalización del tratamiento según el perfil del paciente.
- ¿Qué rol juega el Smart Data en la Industria 4.0?

- Permite la optimización de procesos y la mejora de la calidad.

Videoclase 2. Tecnologías big data

- ¿Qué caracteriza al volumen en el contexto de Big Data?
- La cantidad masiva de datos generados y recopilados.
- ¿Qué implica la velocidad en Big Data?
- La rapidez con la que se generan, recopilan y procesan los datos.
- ¿Qué se entiende por variedad en el contexto de Big Data?
- Los diferentes tipos de datos que se generan y recopilan.
- ¿Por qué es importante la veracidad en Big Data?
- Para asegurar que los datos sean precisos y confiables.
- ¿Qué se busca al extraer valor de los datos en Big Data?
- Transformar datos brutos en insights accionables.

Videoclase 3. Origen de las tecnologías big data

- ¿Cuál es una de las características clave de los datacenters utilizados para Big Data?
- Escalabilidad
- ¿Qué mide el término FLOPS en el contexto del procesamiento de Big Data?
- Las operaciones de punto flotante por segundo.
- ¿Cuál es una diferencia clave entre la computación paralela y la computación distribuida?
- La computación paralela se centra en la ejecución simultánea de tareas dentro de una máquina
- ¿Cuál es el enfoque principal de High-Throughput Computing (HTC)?

- Realizar cálculos científicos complejos. => HTC (High-Throughput Computing) se enfoca en la capacidad de manejar y procesar un gran número de trabajos computacionales independientes en paralelo.
- ¿Cuál es una ventaja de implementar soluciones de Big Data en la nube?
- Escalabilidad fácil y flexible.

Tema 2. HDFS y MapReduce

1. ¿Cuánto ocupa en total un archivo de 500 MB almacenado en HDFS, sin replicación, si se asume el tamaño de bloque por defecto? A. Ocupará 500 MB. -> **correcto** B. Ocupará 512 MB, que son 4 bloques de 128 MB, y hay 12 MB desperdiciados. C. Ocupará lo que resulte de multiplicar 500 MB por el número de datanodes del clúster.
2. ¿Cuál de las siguientes afirmaciones respecto a HDFS es cierta? A. El tamaño de bloque debe ser siempre pequeño para no desperdiciar espacio. B. El factor de replicación es configurable por fichero y su valor, por defecto, es 3. -> **correcto** C. Las dos respuestas anteriores son correctas.
3. ¿Qué afirmación es cierta sobre el proceso de escritura en HDFS? A. El cliente manda al namenode el fichero, que, a su vez, se encarga de escribirlo en los diferentes datanodes. B. El cliente escribe los bloques en todos los datanodes que le ha especificado el namenode. -> **correcto** C. El cliente escribe los bloques en un datanode y este envía la orden de escritura a los demás.
4. En un clúster de varios nodos donde no hemos configurado la topología… A. Es imposible que dos réplicas del mismo bloque caigan en el mismo nodo. B. Es imposible que dos réplicas del mismo bloque caigan en el mismo rack. -> **correcto** C. Las dos respuestas anteriores son falsas.
5. Cuando usamos namenodes federados… A. Cada datanode puede albergar datos de uno de los subárboles. -> **correcto** B. La caída de un namenode no tiene ningún efecto en el clúster. C. Ninguna de las respuestas anteriores es correcta.
6. ¿Por qué se dice que HDFS es un sistema escalable? A. Porque reemplazar un nodo por otro más potente no afecta a los namenodes. B. Porque un clúster es capaz de almacenar datos a gran escala. C. Porque se puede aumentar la capacidad del clúster añadiendo más nodos. -> **correcto**

7. ¿Qué tipo de uso suele darse a los ficheros de HDFS? A. Ficheros de cualquier tamaño que se almacenan temporalmente. B. Ficheros de gran tamaño que se crean, no se modifican, y sobre los que se realizan frecuentes lecturas. -> **correcto** C. Ficheros de gran tamaño que suelen modificarse constantemente.

8. La alta disponibilidad de los namenodes de HDFS implica que… A. La caída de un namenode apenas deja sin servicio al sistema de ficheros durante un minuto antes de que otro entre en acción. -> **correcto** B. Es posible escalar los namenodes añadiendo más nodos. C. La caída de un datanode deja sin servicio al sistema durante unos pocos segundos hasta que este es sustituido.

9. El comando de HDFS para moverse a la carpeta /mydata es… A. hdfs dfs –cd /mydata. B. hdfs dfs –ls /mydata. C. No existe ningún comando equivalente en HDFS. -> **correcto**

10. ¿Qué inconveniente presenta MapReduce? A. No es capaz de procesar datos distribuidos cuando son demasiado grandes. B. Entre las fases map y reduce , siempre lleva a cabo escrituras a disco y movimiento de datos entre máquinas. -> **correcto** C. Es una tecnología propietaria y no es código abierto.

Videoclase 1. Hadoop

- ¿Qué componente de Hadoop se encarga de almacenar grandes cantidades de datos distribuidos en diferentes nodos?
- HDFS. => HDFS (hadoop distributed file system), es el sistema de ficheros distribuido de Hadoop que permite almacenar grandes cantidades de datos en la computación distribuida (nodos de un cluster).
- ¿Cuál de los siguientes componentes gestiona los recursos y la ejecución de tareas en un clúster Hadoop?
- YARN => YARN Hadoop es uno de los principales componentes del framework de la herramienta Apache Hadoop. Significa “Yet Another Resource Negotiator” y es el encargado de administrar los recursos que forman el ecosistema de Apache Hadoop.
- ¿Cuál es el nombre del nodo principal responsable de almacenar los metadatos del sistema de archivos en HDFS?
- NameNode => El NameNode (NN) es el maestro o nodo principal del sistema. No se encarga de almacenar los datos en sí, sino de gestionar su acceso y almacenar sus metadatos

- En la arquitectura HDFS, ¿qué componente almacena efectivamente los bloques de datos?
- DataNode => Los DataNodes (DN) se corresponden con los nodos del clúster que almacenan los datos. Se encarga de gestionar el almacenamiento del nodo.
- ¿Qué función realiza el Secondary NameNode en Hadoop?
- Ayuda a evitar la pérdida de datos del NameNode. => Ejecuta los trabajos de MapReduce.

Videoclase 2. HDFS

- ¿Qué rol desempeña el NameNode en HDFS?
- Mantiene y gestiona los metadatos del sistema de archivos. => (NN). Este componente es único nodo que conoce la lista de ficheros y directorios del clúster. El sistema de ficheros no se puede usar sin el NameNode.
- ¿Cuál de las siguientes afirmaciones describe mejor la función de los DataNodes en HDFS?
- Almacenan los bloques reales de datos. => Almacenan los bloques reales de datos, debido a que se asemeja a una tabla de contenidos, en la que se asignan bloques de datos a DataNodes. Debido a esto, necesita menos espacio de disco, pero más recursos computacionales (memoria y CPU) que los DataNodes.
- ¿Qué sucede si el NameNode falla en un clúster Hadoop sin alta disponibilidad configurada?
- El clúster HDFS se detiene y no puede procesar nuevas peticiones. => Para que el cluster HDFS no se detenga, se puede configurar para que exista un NameNode primario activo y un NameNode secundario en espera (o más de uno a partir de Hadoop 3) que actúa como esclavo. Este último toma el control y responde a las peticiones de los clientes si se detecta algún fallo o el nodo primario deja de estar disponible. Estos nodos deben estar sincronizados y tener los mismos metadatos almacenados. Existen dos mecanismos para conseguir esta sincronización.
- ¿Qué mecanismo utiliza HDFS para asegurar la disponibilidad de los datos almacenados en los DataNodes?

- Replicación de datos en múltiples DataNodes. => Replicación de datos en múltiples DataNodes, en Hadoop 2 se introduce el concepto de alta disponibilidad, evitando que exista un único punto de fallo en el sistema.
- ¿Qué componente de HDFS se encarga de notificar al NameNode sobre la salud y la disponibilidad de los DataNodes?
- Heartbeat. => Para garantizar la tolerancia a fallos, si un datanode deja de estar disponible, el namenode lo detecta mediante un proceso de heartbeat y vuelve a replicar los bloques perdidos en otras máquinas que sí estén disponibles.

Videoclase 3. MAP- Reduce

- En la arquitectura de MapReduce, ¿qué hace la fase de «Map»?
- Divide los datos de entrada en pares clave-valor intermedios. => Divide los datos de entrada en pares clave-valor intermedios, cada nodo esclavo (worker) aplica la función map a los datos locales, y escribe la salida en un almacenamiento temporal. Un nodo maestro garantiza que sólo se procese una copia de los datos de entrada redundantes.
- ¿Cuál es el propósito de la fase de «Reduce» en un trabajo de MapReduce?
- Combinar y procesar los pares clave-valor intermedios generados por los mappers. => los nodos trabajadores procesan cada grupo de datos de salida, por clave, en paralelo.
- ¿Qué componente de Hadoop gestiona la ejecución de trabajos de MapReduce?
- JobTracker. => obTracker (Nota: En versiones más recientes de Hadoop, YARN gestiona los recursos y el JobTracker ha sido reemplazado por el ResourceManager y el ApplicationMaster)
- ¿Qué es un «Split» en el contexto de MapReduce?
- Una división lógica de los datos de entrada que se procesa por un solo mapper. => Una división lógica de los datos de entrada que se procesa por un solo mapper, son los datos que van a ser procesados por el Mapper. Se corresponden con un un bloque del fichero.
- ¿Qué sucede si un DataNode falla durante la ejecución de un trabajo de MapReduce?

- El NameNode reasigna las tareas a otros DataNodes. => El NameNode reasigna las tareas a otros DataNodes. De forma predeterminada, los DataNodes están configurados para escribir en discos de una manera circular. Si se utilizan discos de diferentes capacidades, o si los discos fallan y se reemplazan, el algoritmo de escritura de turnos continuos continúa escribiendo en cada disco a la vez, lo que resulta en un porcentaje diferente de capacidad utilizada en cada disco.

Tema 3. Spark I

1. ¿Cuál es la principal fortaleza de Spark? A. Opera en memoria principal, lo que hace que los cálculos sean mucho más rápidos. -> **correcto** B. Nunca da lugar a movimiento de datos entre máquinas (shuffle). C. Las respuestas A y B son correctas. D. Las respuestas A y B son incorrectas.

2. ¿Qué tipo de procesos se benefician especialmente de Spark? A. Los procesos en modo batch, como, por ejemplo, una consulta SQL. B. Los proceso aplicados a datos no demasiado grandes. C. Los algoritmos de aprendizaje automático que dan varias pasadas sobre los mismos datos. -> **correcto** D. Las respuestas A, B y C son correctas.

3. ¿Cuál es la estructura de datos fundamental en Spark? A. RDD. -> **correcto** B. DataFrame. C. SparkSession. D. SparkContext

4. En una operación de Spark en la que sea necesario movimiento de datos… A. Siempre debemos escribirlos primero en el disco local del nodo emisor. B. No hay acceso al disco local, puesto que Spark opera siempre en memoria. C. Spark nunca provoca movimiento de datos, a diferencia de MapReduce. D. Las respuestas A, B y C son incorrectas. -> **correcto**

5. Elige la respuesta correcta: Cuando se ejecuta una transformación en Spark sobre un RDD… A. Se crea inmediatamente un RDD con el resultado de la transformación. B. Se modifica inmediatamente el RDD con el resultado de la transformación. C. Se añade la transformación al DAG, que creará un RDD con el resultado de la transformación cuando se materialice el RDD resultante. -> **correcto** D. Se añade la transformación al DAG, que modificará el RDD original con el resultado de la transformación cuando se materialice el RDD resultante.

6. Elige la respuesta correcta: La acción collect de Spark… A. No existe como acción; es una transformación. B. Aplica una función a cada fila del RDD de entrada y devuelve otro RDD. C. Lleva todo el contenido del RDD al driver y podría provocar una excepción. -> **correcto** D. Lleva algunos registros del RDD al driver.

7. Elige la respuesta incorrecta: Un PairRDD... A. Es un tipo de RDD que permite realizar tareas de agregación y joins. B. Es un tipo de RDD que contiene una tupla con un número variable de componentes. -> **correcto** C. Es un tipo de RDD cuyo primer componente se considera la clave y el segundo, el valor. D. Se define como cualquier otro RDD, pero con un formato concreto.

8. ¿Qué es un executor de Spark? A. Cada uno de los nodos del clúster de Spark. B. Un proceso creado en los nodos del clúster, preparado para recibir trabajos de Spark. -> **correcto** C. Un nodo concreto del clúster que orquesta los trabajos ejecutados en él. D. Ninguna de las definiciones anteriores es correcta.

9. La acción map de Spark... A. No existe como acción; es una transformación. -> **correcto** B. Aplica una función a cada fila del RDD de entrada y devuelve otro RDD. C. Lleva todo el contenido del RDD al driver y podría provocar una excepción. D. Lleva ciertos registros del RDD al driver.

10. Cuando Spark ejecuta una acción... A. Se materializan en la memoria RAM de los workers todos los RDD intermedios necesarios para calcular el resultado de la acción y después se liberan todos. B. Se añade la acción al DAG y no hace nada en ese momento. C. Se materializan los RDD intermedios necesarios que no estuviesen ya materializados, se calcula el resultado de la acción y se liberan los no cacheados. -> **correcto** D. Ninguna de las respuestas anteriores es correcta.

Videoclase 1. Spark

- ¿Cuál es la principal ventaja de Apache Spark sobre Hadoop MapReduce?
- Procesamiento de datos en memoria, lo que aumenta la velocidad.
- ¿Qué componente de Apache Spark es responsable de gestionar la distribución de tareas y recursos en un clúster?
- ClusterManager => El administrador de clúster o Clúster Manager en Spark hace referencia a la comunicación del driver con el backend para adquirir recursos físicos y poder ejecutar los executors.
- ¿Cuál de los siguientes lenguajes de programación no es soportado de manera nativa por Apache Spark?
- Ruby.
- ¿Qué es un RDD (Resilient Distributed Dataset) en Apache Spark?

- Una estructura de datos fundamental que permite el procesamiento distribuido tolerante a fallos.
- ¿Qué operación en Apache Spark se utiliza para aplicar una función a cada elemento de un RDD y devolver un nuevo RDD?
- Map => Esta función se utiliza para aplicar una transformación a cada elemento de un RDD (Resilient Distributed Dataset). Toma una función como argumento y aplica esa función a cada elemento del RDD, devolviendo un nuevo RDD con los resultados.

Videoclase 2. Spark Arquitectura

- ¿Cuál es la estructura de datos fundamental en Apache Spark que permite el procesamiento distribuido y tolerante a fallos?
- RDD (Resilient Distributed Dataset). => RDD (Resilient Distributed Dataset): Es la principal abstracción de datos, el tipo de dato básico que tiene Apache Spark. Los RDD están particionados en los distintos nodos del clúster, ya que Apache Spark se suele instalar en un clúster o conjunto de máquinas, por lo que esos RRDs se encuentran distribuidos sobre esas máquinas. Con ello se consigue la tolerancia a fallos, porque si falla una máquina tenemos el fichero en otras máquinas.
- ¿Qué componente de Apache Spark gestiona los recursos del clúster y asigna recursos para las aplicaciones?
- ClusterManager. => ClusterManager: El administrador de clúster o Clúster Manager en Spark hace referencia a la comunicación del driver con el backend para adquirir recursos físicos y poder ejecutar los executors.
- ¿Cuál es la diferencia principal entre un DataFrame y un RDD en Apache Spark?
- Los DataFrames proporcionan una API optimizada y utilizan un planificador de consultas, mientras que los RDDs son colecciones distribuidas de objetos. => Tanto los RDD como los conjuntos de datos proporcionan una API de estilo OOP, mientras que los DataFrames proporcionan una API de estilo SQL. En los RDD, le especificamos al motor Spark cómo realizar una determinada tarea, mientras que, con los DataFrames y los conjuntos de datos, especificamos qué hacer y el motor Spark se encarga del resto.
- ¿En qué modo de despliegue de Apache Spark el Driver Program corre en el mismo proceso que la aplicación?

- Local Mode.=>En modo local toda la infraestructura de Spark se aloja en una sola JVM dentro de una sola computadora, el driver y el resource manager tambien se encuentran alojados.
- ¿Qué componente en Apache Spark es responsable de ejecutar una serie de transformaciones y acciones en un RDD?
- Executor. => Executor: Los executors en Spark hacen referencia al proceso en el que estos realizan la carga de trabajo. De manera que los executors obtienen sus tareas desde el driver y llevan a cabo la carga, la transformación y el almacenamiento de los datos.

Videoclase 3. Spark Transformaciones

- ¿Cuál es la característica principal de las transformaciones en Spark?
- Son perezosas y construyen un plan de ejecución que se ejecuta cuando se llama a una acción. => Transformaciones Lazy: Las transformaciones en RDD son «lazy» (perezosas), lo que significa que no se ejecutan de inmediato. En su lugar, se registran y se ejecutan solamente cuando se realiza una acción. Esto permite la optimización y la ejecución eficiente de las operaciones.
- ¿Qué transformación en Apache Spark devuelve un nuevo RDD que solo contiene los elementos del RDD original que satisfacen una función predicada?
- Filter => En Spark, la función Filtro devuelve un nuevo conjunto de datos formado al seleccionar aquellos elementos de la fuente en los que la función devuelve verdadero. Por lo tanto, recupera sólo los elementos que satisfacen la condición dada..
- ¿Cuál transformación en Apache Spark combina múltiples RDDs en un solo RDD?
- Union => Nos devuelve la unión de dos o más RDDs
- ¿Qué transformación en Apache Spark agrupa los elementos de un RDD según una clave y devuelve un RDD de pares clave-valor?
- GroupByKey => Agrupa los valores de cada clave en el RDD en una única secuencia Hash particiona el RDD resultante con particiones numPartitions.
- ¿Qué transformación en Apache Spark une dos RDDs de pares clave-valor por sus claves y devuelve un nuevo RDD de pares clave y tuplas de valores?

- Join => Devuelve un RDD que contiene todos los pares de elementos con claves coincidentes en self y other. Cada par de elementos se devolverá como una tupla (k, (v1, v2)), donde (k, v1) está en uno mismo y (k, v2) está en otro.Realiza una unión hash en todo el clúster.

TEMA 4: Spark II

1. Elige la respuesta correcta respecto a los DataFrames de Spark: A. Un RDD es una envoltura de un DataFrame de objetos de tipo Row. B. Un DataFrame es una envoltura de un RDD de objetos de tipo Row. -> **correcto** C. Un DataFrame es una envoltura de un objeto de tipo Row que contiene RDD. D. Ninguna de las respuestas anteriores es correcta.

2. Elige la respuesta correcta sobre los DataFrames de Spark: A. Puesto que representan una estructura de datos más compleja que un RDD, no es posible distribuirlos en memoria. B. Puesto que son un envoltorio de un RDD, suponen una estructura de datos que sigue estando distribuida en memoria. -> **correcto** C. Son una estructura de datos no distribuida en memoria, al igual que los DataFrames de Python o los data.frames de R. D. Son una estructura de datos distribuida en disco.

3. ¿Qué mecanismo ofrece la API estructurada de DataFrames para leer datos? A. Método read de la Spark Session. -> **correcto** B. Método read del Spark Context. C. No ofrece ningún método, sino que se utiliza la API de RDD para leer datos. D. Método ingest de la Spark Session.

4. ¿Es obligatorio especificar explícitamente el esquema del DataFrame cuando se leen datos de fichero? A. No, porque solo se pueden leer ficheros estructurados como Parquet, que ya contienen información sobre su esquema. B. Sí, porque, si no se indica el esquema, Spark no es capaz de leer ficheros CSV, ya que no sabe con qué tipo almacenar cada campo. C. No, porque, si no se indica el esquema, Spark guardará todos los campos de los que no sepa su tipo como strings. -> **correcto** D. No, porque, si no se indica el esquema y se intenta leer ficheros sin esquema implícito, Spark lanzará un error.

5. Seleccione la respuesta incorrecta: ¿Por qué es aconsejable utilizar DataFrames en Spark en lugar de RDD? A. Porque son más intuitivos y fáciles de manejar a alto nivel. B. Porque son más rápidos, debido a optimizaciones realizadas por Catalyst. C. Porque los DataFrames ocupan menos en disco. -> **correcto** D. Las respuestas A y B son correctas.

6. Tras ejecutar la operación b = df.withColumn("nueva", 2*col("calif")): A. El DataFrame contenido en df tendrá una nueva columna, llamada nueva. B. Llevaremos al driver el resultado de multiplicar 2 por la columna calif. C. El DataFrame contenido en b tendrá una columna más que df. -> **correcto** D. El DataFrame contenido en b tendrá una única columna llamada nueva

7. ¿Cuál es la operación con la que nos quedamos con el subconjunto de filas de un DataFrame que cumplen una determinada condición? A. sample. B. filter. -> **correcto** C. map. D. show.

8. Las API estructuradas de DataFrames y Spark SQL... A. Son API que no se pueden combinar: una vez que se empieza a usar una de ellas, se tienen que hacer todas las tareas con la misma API. B. Se pueden aplicar funciones de la API de DataFrames sobre el resultado de consultas de Spark SQL. -> **correcto** C. Se pueden aplicar el método s q l para lanzar consultas SQL sobre DataFrames sin registrar. D. Ninguna de las opciones anteriores es correcta.

9. La transformación map de Spark... A. No se puede aplicar a un DataFrame porque pertenece a la API de RDD. B. Se puede aplicar a un DataFrame porque pertenece a la API estructurada de DataFrames. C. Se puede aplicar a un DataFrame porque envuelve un RDD al que se puede acceder mediante el atributo rdd . -> **correcto** D. No existe en Spark; map es una acción

Videoclase 1. Spark - Dataframes

- ¿Qué es un DataFrame en Spark?
- Una colección distribuida de datos organizados en columnas nombradas. => Un DataFrame en Spark es una colección distribuida de datos organizados en columnas nombradas, similar a una tabla en una base de datos relacional.
- ¿Cuál de las siguientes es una fuente de datos válida para crear un DataFrame en Spark?
- Archivos CSV
- ¿Cuál es el método correcto para leer un archivo CSV en un DataFrame en Spark?
- spark.read.csv("ruta/al/archivo.csv")
- ¿Cuál es el propósito de SparkSession en Apache Spark?

- Crear DataFrames y ejecutar SQL sobre ellos. => SparkSession es el punto de entrada para todas las funcionalidades de Spark, permitiendo crear DataFrames, ejecutar SQL, crear vistas temporales y leer datos de diversas fuente
- ¿Qué implica definir un esquema en el contexto de Spark?
- Definir la estructura de los datos con columnas y tipos de datos. => Definir un esquema en Spark implica especificar la estructura de los datos, incluyendo las columnas y sus tipos de datos, lo que facilita su manipulación y análisis.

NOTA

- El analisis EDA en resumen es la visualición de los calculos de los datos del dataset debemos validar su:
  - count() Total de datos
  - mean() promedio
  - stddev() desviación standar => indica qué tan dispersos están los datos con respecto a su media. En otras palabras, muestra la variabilidad o dispersión de un conjunto de datos
  - min() mínimos
  - Max() máximos

Videoclase 2. Manipulación de dataframes

- ¿Qué significa que los dataframes en Spark sean inmutables?
- Las operaciones sobre dataframes devuelven un nuevo dataframe sin modificar el original. => La inmutabilidad de los dataframes en Spark significa que cualquier operación realizada sobre ellos devuelve un nuevo dataframe, dejando el dataframe original sin cambios.
- ¿Cuál es la diferencia entre transformaciones y acciones en Spark?
- Las transformaciones devuelven un nuevo dataframe y las acciones desencadenan la ejecución de las transformaciones. => Las transformaciones en Spark devuelven un nuevo dataframe sin ejecutar la operación hasta que se realiza una acción, que es la que desencadena la ejecución de todas las transformaciones previas.
- ¿Qué hace el método count() cuando se aplica a un dataframe en Spark?
- Cuenta las filas del dataframe y devuelve un entero. => El método count() cuenta las filas del dataframe y devuelve el resultado como un entero.

- ¿Qué realiza la operación collect() en Spark cuando se aplica a un dataframe?
- Devuelve todas las filas del dataframe al driver como una lista de objetos row. => La operación collect() devuelve todas las filas del dataframe al driver como una lista de objetos row, lo que implica que el resultado debe caber en la memoria del driver.
- ¿Qué hace la operación take No cuando se aplica a un dataframe en Spark?
- Devuelve las primeras n filas del dataframe como una lista de objetos row.

Videoclase 3. Spark SQL

- ¿Qué hace la transformación select en un dataframe de Spark?
- Selecciona y devuelve un nuevo dataframe que solo contiene las columnas especificadas
- ¿Cuál es el propósito de la transformación withColumn en Spark SQL?
- Crear una nueva columna o reemplazar una existente en el dataframe.
- ¿Qué hace la transformación filter en un dataframe de Spark?
- Selecciona filas del dataframe que cumplen con una condición.
- ¿Cuál es el resultado de utilizar groupBy seguido de agg en un dataframe de Spark?
- Agrupa los datos por una o más columnas y aplica agregaciones. => La operación groupBy seguida de agg agrupa los datos por una o más columnas y permite aplicar funciones de agregación a cada grupo.
- ¿Cuál es el propósito del método sql en Spark SQL?
- Ejecutar consultas SQL sobre dataframes registrados como tablas. => El método sql en Spark SQL permite ejecutar consultas SQL sobre dataframes que han sido registrados como tablas, traduciendo las consultas a trabajos optimizados de Spark.

Tema 5. Spark III

1. ¿Qué diferencia Spark MLlib de Spark ML? A. Spark MLlib ofrece interfaz para DataFrames en todos sus componentes, mientras que Spark ML sigue utilizando RDD y ha quedado obsoleta. B. Spark MLlib no permite cachear los resultados de los modelos, mientras que Spark ML sí. C. Spark MLlib es más rápida entrenando modelos que Spark ML. D. Ninguna de las respuestas anteriores es correcta. -> **correcto**

2. ¿Qué tipo de componentes ofrece Spark ML? A. Estimadores y transformadores para ingeniería de variables y para normalizar datos. B. Estimadores y transformadores para preparar los datos para el formato requerido por los algoritmos de aprendizaje automático de Spark. C. Solo pipelines que no dan acceso a los estimadores internos. D. Las respuestas A y B anteriores son correctas. -> **correcto**

3. ¿Cuál es el método principal de un estimator de Spark ML? A. El método fit. -> **correcto** B. El método transform. C. El método estimate. D. El método describe.

4. ¿A qué interfaz pertenecen los algoritmos de machine learning de Spark cuando aún no han sido entrenados? A. Transformer. B. Estimator. -> **correcto** C. Pipeline. D. DataFrame.

5. ¿A qué interfaz pertenecen los modelos de Spark ML cuando ya han sido entrenados con datos? A. Transformer. -> **correcto** B. Estimator. C. Pipeline. D. DataFrame.

6. ¿Qué ocurre si creamos un StringIndexer para codificar las etiquetas de una variable en el dataset de entrenamiento y después creamos otro StringIndexer para codificar los datos de test en el momento de elaborar predicciones? A. Obtendremos la misma codificación en los dos. B. Da un error, porque un mismo StringIndexer no puede añadirse a dos pipelines. C. Podríamos obtener codificaciones distintas de la misma etiqueta en los datos de entrenamiento y en los de test, lo que falsearía los resultados de las predicciones. -> **correcto** D. Ninguna de las respuestas anteriores es correcta.

7. ¿Cuál es la estructura principal que maneja Spark Structured Streaming? A. DStreams. B. DStreams DataFrames. C. Streaming DataFrames. -> **correcto** D. Streaming RDD.

8. Spark Streaming permite leer flujos de datos: A. Solo desde tecnologías de ingesta de datos como Apache Kafka. B. Desde cualquier fuente de datos, siempre que contenga un esquema, como, por ejemplo, una base de datos. C. Desde fuentes como Apache Kafka y HDFS, si activamos la inferencia de esquema. D. Las respuestas A, B y C son incorrectas.

9. En Spark Streaming, una vez se ejecuta la acción start: A. El driver espera automáticamente a que concluya la recepción de flujo para finalizar su ejecución. B. Hay que ejecutar un método para indicar al driver que no finalice automáticamente y que espere a que concluya la recepción del flujo. C. Un flujo de datos no tiene fin y, por tanto, el driver nunca puede finalizar. D. Ninguna de las opciones anteriores es correcta. -> **correcto**

10. ¿Qué acciones pueden realizarse en Spark Structured Streaming? A. take. B. show. C. start. -> **correcto** D. collect.

Videoclase 1. Spark MLlib

- ¿Qué es Spark MLlib?
- Una biblioteca de aprendizaje automático de Spark. => Es la biblioteca de aprendizaje automático de Apache Spark que proporciona herramientas para el análisis de datos, incluyendo algoritmos de clasificación, regresión, clustering, filtrado colaborativo, y más.
- ¿Cuál de los siguientes es un algoritmo de clasificación disponible en Spark MLlib?
- Random Forest.=> es un algoritmo de clasificación y regresión que está disponible en Spark MLlib. K-means es un algoritmo de clustering y Linear Regression es un algoritmo de regresión.
- ¿Qué método se utiliza para dividir un DataFrame en conjuntos de entrenamiento y prueba en Spark MLlib?
- RamdonSplit() => El método `randomSplit()` se utiliza para dividir un DataFrame en varios subconjuntos de forma aleatoria. Por ejemplo, `df.randomSplit([0.8, 0.2])` divide el DataFrame en dos conjuntos, uno para entrenamiento (80%) y otro para prueba (20%).
- ¿Cuál de los siguientes es un algoritmo de clustering disponible en Spark MLlib?
- K-means. => es un algoritmo de clustering que está disponible en Spark MLlib. Decision Tree y Logistic Regression son algoritmos de clasificación.
- ¿Cómo se puede evaluar el rendimiento de un modelo de clasificación en Spark MLlib?
- Usando la métrica de Silhouette score.

NOTA: TENER CLARO ESTO

- K-means. => es un algoritmo de clustering que está disponible en Spark MLlib.
- Decision Tree y Logistic Regression son algoritmos de clasificación.
- Random Forest.=> es un algoritmo de clasificación y regresión que está disponible en Spark MLlib.
- K-means es un algoritmo de clustering
- Linear Regression es un algoritmo de regresión.

Videoclase 2. Spark Streaming y structured streaming

- ¿Qué es Spark Streaming?
- Una biblioteca de Spark para el procesamiento y análisis de datos en tiempo real. => es una extensión de Apache Spark que permite el procesamiento de flujos de datos en tiempo real. Se usa para procesar datos en tiempo real provenientes de diversas fuentes como Kafka, Flume, Kinesis, o sockets TCP.
- ¿Qué objeto en Spark Streaming se utiliza para definir un flujo de datos?
- DStream. => Un `DStream` (Discretized Stream) es la abstracción básica en Spark Streaming que representa un flujo continuo de datos. Cada `DStream` es una secuencia continua de RDDs.
- ¿Qué método se usa para crear un contexto de streaming en Spark Streaming?
- StreamingContext() => `StreamingContext` es el objeto principal en Spark Streaming que se utiliza para crear un contexto de streaming. Ejemplo: `ssc = StreamingContext(sparkContext, batchDuration)` .
- ¿Cómo se puede leer un flujo de datos desde un socket TCP en Spark Streaming?
- ssc.socketTextStream(hostname, port) => El método `socketTextStream()` se utiliza para leer un flujo de datos de texto desde un socket TCP. Ejemplo: `ssc.socketTextStream("localhost", 9999)` .
- ¿Cómo se puede iniciar el procesamiento de un flujo de datos en Spark Streaming?
- ssc.start() => se utiliza para iniciar el procesamiento de datos en el contexto de streaming. Ejemplo: `ssc.start()` .

Videoclase 3. Ejemplos de MLlib y Streaming

- ¿Cuál es el propósito de la opción outputMode en Spark Structured Streaming?

- Especificar cómo se estructuran los resultados del flujo de salida. => La opción outputMode especifica cómo se estructuran los resultados del flujo de salida. Los modos de salida comunes son `append`, `complete`, y `update`. Ejemplo: `df.writeStream.outputMode("append").format("console").start()`.
- ¿Qué es un trigger en Spark Structured Streaming?
- es una configuración que controla la frecuencia con la que se procesa el flujo de datos. Ejemplo: `df.writeStream.trigger(processingTime='10 seconds').start()`.
- ¿Cómo se puede especificar una fuente de datos Kafka en Spark Structured Streaming?
- spark.readStream.kafka("topics") => Para especificar una fuente de datos Kafka en Spark Structured Streaming, se utiliza el método `readStream` con el formato `kafka` y la opción `subscribe`. Ejemplo: `spark.readStream.format("kafka").option("subscribe", "topic_name").load()`.
- ¿Cuál de los siguientes modos de salida en Spark Structured Streaming solo agrega nuevas filas al resultado?
- append. => El modo de salida `append` solo agrega nuevas filas al resultado, sin modificar las filas existentes. Es adecuado para flujos donde solo se necesita agregar datos nuevos. Ejemplo: `df.writeStream.outputMode("append").format("console").start()`.
- ¿Qué método se usa para aplicar una transformación de ventana en un flujo de datos en Spark Structured Streaming?
- groupBy() => el método `window()` se utiliza para aplicar una transformación de ventana en un flujo de datos, permitiendo agrupar datos en intervalos de tiempo. Ejemplo: `df.groupBy(window("timestamp", "10 minutes")).count()`.

-
-
-
-

Tema 6. Apache Kafka

1. ¿Qué es Apache Kafka? A. Un sistema de mensajería que utiliza Spark para funcionar. B. Un bus de datos distribuido, en el que varias aplicaciones pueden leer y escribir. -> **correcto** C. Un sistema de colas basado en MapReduce. D. Un sistema de data warehousing.

2. Cuando usamos Kafka... A. Cada aplicación elige el tipo de mensajes que desea leer. -> **correcto** B. Todas las aplicaciones reciben todos los mensajes. C. Solo las aplicaciones registradas en Spark pueden acceder al bus. D. Cada aplicación puede leer solo un tipo de mensajes.

3. ¿Cuál de estas funciones es típica de Kafka? A. Transmitir mensajes generados por una aplicación a otras que los utilizan. -> **correcto** B. Almacenar información accesible para distintas aplicaciones, tal como lo hace una base de datos. C. Realizar procesados de flujos de información. D. Ninguna de las opciones anteriores es correcta.

4. Un topic de Kafka... A. Es una partición de los datos subyacentes. -> **correcto** B. Es un conjunto de mensajes que comparten la misma estructura. C. Es equivalente a una base de datos. D. Es una máquina que almacena cierto tipo de datos.

5. Si en un grupo de consumidores hay más consumidores suscritos a un topic que particiones tiene dicho topic: A. Kafka reparte los mensajes entre todos consumidores de una misma partición. B. Kafka no permite que esto ocurra y denegará la suscripción al consumidor. C. Uno o más consumidores quedarán ociosos, sin poder consumir mensajes. -> **correcto** D. Todos los consumidores reciben mensajes de todas las particiones.

6. Cuando un proceso productor de Kafka utiliza envío asíncrono: A. Se bloquea en espera de la respuesta que confirme que todo ha ido bien. B. Prosigue su ejecución, ya que, al ser asíncrono, no espera respuesta alguna. C. Prosigue su ejecución y Kafka invocará el método que el productor indicó cuando tenga disponible la respuesta. -> **correcto** D. Ninguna de las respuestas anteriores es correcta.

7. ¿Qué implica que un bróker contenga la partición líder de un topic? A. Que será quien reciba y procese las peticiones de lectura y escritura a esapartición. -> **correcto** B. Que decidirá si un consumidor está autorizado para suscribirse al topic. C. Que será quien centralice las peticiones de escritura que reciben todos losbrókeres que contengan dicha partición. D. Las tres opciones anteriores son correctas.

8. ¿Cuál de las siguientes afirmaciones sobre Kafka es correcta? A. Un mensaje en Kafka se elimina según el primer consumidor lo lee. B. Un mensaje en Kafka se elimina siempre tras pasar una semana en elbróker. C. Un mensaje en Kafka se elimina cuando lo han leído el númeroconfigurado de consumidores. D. Un mensaje en Kafka se elimina cuando se cumple el tiempo o tamañoconfigurados. -> **correcto**

9. El bróker encargado de supervisar qué brókeres se unen y cuáles dejan el clústeres: A. El bróker líder. B. El bróker controlador. -> **correcto** C. El bróker sincronizado (in-sync). D. Se encarga Zookeeper.

10. Un clúster Kafka está compuesto: A. Por uno o varios brókeres, y siempre incluye Zookeeper para gestionar metadatos. B. Por, al menos, dos o más brókeres, y se complementa a veces con Zookeeper para facilitar la gestión de metadatos. C. Por uno o varios brókeres, y se complementa a veces con Zookeeper para facilitar la gestión de metadatos. -> **correcto** D. Ninguna de las opciones es correct

Videoclase 1. Apache Kafka: Mensajeria publicación/suscripción

- ¿Qué es Apache Kafka?
- Un sistema de mensajería basado en publicación-suscripción. => se describe como un bus de datos único de métricas basado en el patrón de publicación-suscripción
- ¿Cuál es una característica principal del patrón de publicación-suscripción utilizado por Apache Kafka?
- Los receptores se suscriben a clases de mensajes para recibir los datos. => En el patrón de publicación-suscripción de Apache Kafka, los emisores clasifican los mensajes bajo ciertas clases (topics), y los receptores interesados se suscriben a estas clases para recibir los mensajes

- ¿Cómo se asegura Apache Kafka de manejar el gran volumen de datos en aplicaciones de e-commerce? -Utilizando múltiples servidores frontend para recopilar métricas. => En una aplicación web de compras, se utilizan varios servidores frontend para recopilar diferentes métricas de los usuarios y enviar estos datos a un servidor encargado de procesarlos
- ¿Cuál es el propósito de utilizar Apache Kafka en una gran aplicación web de compras?
- Mejorar la experiencia del usuario en tiempo real mediante recomendaciones.
- ¿Cuál de las siguientes opciones no es una función de Apache Kafka?
- Procesamiento en tiempo real de transacciones financieras. => se utiliza principalmente para la publicación y suscripción de mensajes y para el transporte de métricas de navegación, pero no específicamente para el procesamiento en tiempo real de transacciones financieras

Videoclase 2. Apache Kafka: conceptos fundamentales

- ¿Qué es Apache Kafka?
- Un bus de datos distribuido y replicado. => Apache Kafka es un bus de datos (también llamado cola de mensajes) distribuido y replicado basado en el paradigma publicación/suscripción. Se utiliza para la mensajería entre aplicaciones, donde los mensajes se insertan y consumen en un cierto orden
- ¿Cuál es la unidad de información en Kafka?
- Mensaje. => En Kafka, la unidad de información es el mensaje, que es equivalente a un registro (fila) de una base de datos. Un mensaje es simplemente un array de bytes sin significado para Kafka
- ¿Qué es un topic en Kafka?
- Una tabla de base de datos => En Kafka, un topic es equivalente a una tabla de base de datos e indica la agrupación de los mensajes, estructurando todos de la misma forma. Los topics contienen particiones para escalabilidad y replicación de los datos
- ¿Cuál es el papel de Zookeeper en un clúster de Kafka?

- Almacenar metadatos del clúster. => Zookeeper almacena los metadatos del clúster de Kafka, incluyendo una lista de los brokers del clúster. Cada bróker tiene un Id único para registrarse en Zookeeper, y el primer bróker que se registra actúa como controlador que selecciona las particiones líderes.

- ¿Cómo se sincronizan los brokers con réplicas followers en Kafka?

- Solicitando regularmente los últimos mensajes al bróker líder. => Para mantenerse sincronizados, los brokers con réplicas followers solicitan regularmente al bróker con la réplica líder los últimos mensajes que este haya recibido. El líder es responsable de saber qué réplicas followers están sincronizadas y cuáles no

Videoclase 3. Productores y consumidores

- ¿Cómo se sincronizan los brokers con réplicas followers en Kafka?

- Serializa la clave y el valor del mensaje. => Kafka primero serializa la clave y el valor del mensaje, convirtiéndolos en secuencias binarias de bytes que pueden ser transmitidos por la red

- ¿Cuál de los siguientes no es un método de envío en Kafka?

- Envío encriptado. => Kafka soporta tres métodos de envío: enviar y olvidar, envío síncrono y envío asíncrono. El envío encriptado no es un método soportado por Kafka.

- ¿Qué ocurre cuando el broker recibe un bloque de mensajes en Kafka?

- Envía una respuesta si el bloque se escribió con éxito. => Cuando el broker recibe un bloque de mensajes, envía una respuesta indicando si la escritura fue exitosa. Si se escribió con éxito, la respuesta es un objeto RecordMetadata, de lo contrario, se puede intentar reenviar antes de devolver un error.

- ¿Qué método de envío en Kafka permite al productor continuar su ejecución sin esperar la respuesta?

- Envío asíncrono. => En el envío asíncrono, el productor incluye un callback que Kafka invoca automáticamente cuando se recibe la respuesta, permitiendo al proceso continuar su ejecución sin esperar la respuesta.

- ¿Cuál es la función de un callback en el envío asíncrono en Kafka?

- Notificar al productor sobre el éxito o fracaso del envío. => En el envío asíncrono, el callback es invocado automáticamente por Kafka cuando se recibe la respuesta y permite examinar el éxito o fracaso del envío.

Tema 7. Hive e Impala

1. Un ejemplo ideal de alguien que puede utilizar Hive es: A. Un analista con conocimientos de SQL que quiere consultar datos estructurados almacenados en HDFS. -> **correcto** B. Un programador con conocimientos de MapReduce que quiere consultar imágenes y vídeos. C. Una persona de negocios, con alto conocimiento de Excel, que quiere consultar rápidamente datos masivos guardados en una base de datos relacional como MySQL. D. Los tres casos anteriores son buenos casos de uso.

2. Sobre Apache Hive: A. Existen versiones libres y de pago. B. Permite consultar archivos almacenados en HDFS utilizando lenguaje SQL. -> **correcto** C. Requiere poseer una base de datos relacional que funcione como respaldo. D. Solo se puede usar como parte de la distribución de Cloudera.

3. Hive se define como: A. Una base de datos SQL distribuida. B. Un motor de ejecución distribuido para consultas SQL. C. Una base de datos NoSQL distribuida. D. Un traductor de consultas SQL a trabajos de procesado distribuidos. -> **correcto**

4. Para usar Hive: A. Solo se puede utilizar a través de un intérprete de línea de comandos. B. Se puede usar únicamente a través de una conexión JDBC. C. Es posible usarlo desde herramientas de BI que dispongan de conector ODBC. -> **correcto** D. Ninguna de las respuestas anteriores es correcta.

5. ¿Cuál de las siguientes afirmaciones sobre Hive es correcta? A. Hive siempre utiliza como motor de ejecución Apache Spark. B. MySQL puede funcionar como metastore de Hive. -> **correcto** C. Un fichero de texto plano puede funcionar como metastore de Hive. D. Ninguna de las opciones anteriores es correcta.

6. ¿Cuál de las siguientes afirmaciones sobre Hive es correcta? A. Cuando se ejecuta la sentencia DROP sobre una tabla, Hive siempre borra los metadatos relacionados con dicha tabla. B. Cuando se ejecuta la sentencia DROP sobre una tabla, Hive siempre borra los datos relacionados con esta tabla C. Cuando se ejecuta la sentencia DROP sobre una tabla, Hive nunca borra ningún dato ni metadato. D. Cuando se ejecuta la sentencia DROP sobre una tabla, Hive siempre borra los datos y metadatos. -> **correcto**

7. Señala la respuesta correcta: A. Impala está pensado para procesados en bloque (batch), mientras que Hive está dirigido a peticiones interactivas. B. Impala está dirigido a peticiones interactivas, mientras que Hive está pensado para procesados en bloque (batch). -> **correcto** C. Tanto Impala como Hive están pensados para peticiones interactivas. D. Tanto Impala como Hive están pensados para procesados en bloque.

8. ¿Cuál de las siguientes afirmaciones sobre Impala es correcta? A. Impala utiliza como motor de ejecución Apache Spark. B. Impala utiliza como motor de ejecución Apache tez. C. El motor de ejecución de Impala es configurable, igual que en Hive. D. Ninguna de las opciones anteriores es correcta. -> **correcto**

9. El proceso de Impala que se encarga de ejecutar las consultas del usuario es... A. El proceso statestored. B. El proceso impalad. -> **correcto** C. El proceso initd. D. El proceso catalogd.

10. La manera de ejecutar Impala en un clúster de ordenadores es... A. Mediante un proceso que está corriendo en cada máquina y accede directamente a los datos de HDFS de ese nodo. -> **correcto** B. Mediante el motor de ejecución de Apache Spark que se ejecuta en el clúster y sobre el cual nos proporciona una abstracción SQL. C. Mediante las consultas SQL traducidas por Impala al metastore de Hive. D. Ninguna de las anteriores es correcta.

Videoclase 1. Apache Hive

- ¿Qué componente de la arquitectura de Apache Hive almacena los metadatos de las tablas, columnas, tipos de datos y otras estructuras?

- Metastore.=> El Metastore en Apache Hive es el componente que almacena los metadatos de las tablas, columnas, tipos de datos y otras estructuras de datos, permitiendo a Hive gestionar el esquema y las definiciones de los datos

- ¿Cuál es la función del Driver en la arquitectura de Apache Hive?

- Gestionar la sesión de Hive. => El Driver en la arquitectura de Apache Hive actúa como el gestor de la sesión de Hive, procesando las consultas de HiveQL, preparándolas para la compilación y enviándolas al Compiler

- ¿Qué tecnología utiliza Apache Hive para el procesamiento y análisis de grandes volúmenes de datos?

- MapReduce. => Apache Hive utiliza MapReduce como uno de los motores de ejecución para procesar y analizar grandes volúmenes de datos, aprovechando la escalabilidad y la tolerancia a fallos de Hadoop.
- ¿Para qué tipo de análisis de datos se utiliza principalmente Apache Hive?
- OLAP.=> Apache Hive se utiliza principalmente en modo OLAP (Online Analytical Processing) para el análisis en bloque de información empresarial, permitiendo realizar consultas multidimensionales interactivas y análisis de grandes volúmenes de datos
- ¿Qué es una UDF (User-Defined Function) en el contexto de Apache Hive?
- Una función personalizada creada por el usuario. => Una UDF (User-Defined Function) es una función personalizada que los usuarios pueden crear para extender las capacidades de los sistemas de bases de datos o herramientas de análisis de datos, permitiendo escribir funciones específicas que no están disponibles en el lenguaje de consulta SQL estándar.

Videoclase 2. Impala

- ¿Qué es Apache Impala?
- Un motor de consultas SQL de procesamiento masivamente paralelo => Apache Impala es un motor de consultas SQL de código abierto y de procesamiento masivamente paralelo (MPP) diseñado para sistemas de almacenamiento de datos como Hadoop
- ¿Cuál es una de las principales ventajas de Impala en comparación con otros sistemas de procesamiento de Hadoop?
- Consultas de baja latencia y alto rendimiento. => A diferencia de otros sistemas de procesamiento de Hadoop, Impala ofrece consultas de baja latencia y alto rendimiento, lo que lo hace ideal para análisis en tiempo real
- ¿Qué componentes clave forman parte de la arquitectura de Impala?
- Daemon ImpalaD, coordinador de consultas y catálogo de metadatos. => La arquitectura de Impala consta de varios componentes clave como el daemon ImpalaD, el coordinador de consultas y el catálogo de metadatos, que son esenciales para su funcionamiento eficiente
- ¿Cuál es uno de los principales casos de uso de Apache Impala?

- Análisis en tiempo real en sectores como finanzas, salud y comercio electrónico. => Impala se utiliza para realizar análisis en tiempo real en sectores donde la velocidad de las consultas es crítica, como finanzas, salud y comercio electrónico
- ¿Qué permite la arquitectura distribuida de Impala?
- Procesar datos en paralelo a través de múltiples nodos en un clúster de Hadoop. => La arquitectura distribuida de Impala permite procesar datos en paralelo a través de múltiples nodos en un clúster de Hadoop, optimizando el rendimiento y escalabilidad de las consultas.

Videoclase 3. Ejemplos

- ¿Cuál es la función principal de Apache Hive?
- Proporcionar una interfaz SQL para consultar datos almacenados en Hadoop. => Apache Hive es una herramienta de data warehousing construida sobre Hadoop que proporciona una interfaz de consulta similar a SQL, llamada HiveQL, para interactuar con los datos almacenados en HDFS (Hadoop Distributed File System)
- ¿Cuál es una de las principales ventajas de utilizar Apache Impala sobre Apache Hive?
- Velocidad de consulta significativamente más rápida. => Apache Impala permite realizar consultas SQL de baja latencia directamente en datos almacenados en Hadoop, ofreciendo tiempos de respuesta mucho más rápidos en comparación con Hive, que depende del procesamiento en batch de MapReduce.
- ¿Qué tipo de datos puede procesar Apache Hive?
- Datos estructurados y semi-estructurados. => Apache Hive está diseñado principalmente para manejar datos estructurados almacenados en tablas, pero también puede trabajar con datos semi-estructurados como JSON y XML mediante el uso de formatos de almacenamiento como ORC y Parquet.
- ¿Qué comando en Hive se utiliza para crear una nueva tabla?
- CREATE TABLE. => El comando estándar en Hive para la creación de una nueva tabla es `CREATE TABLE`, que permite definir el esquema de la tabla y las opciones de almacenamiento
- ¿Cuál es una diferencia clave entre Apache Hive y Apache Impala en términos de arquitectura de ejecución?

- Hive utiliza MapReduce, mientras que Impala utiliza un motor de ejecución dedicado. => Apache Hive traduce consultas SQL a trabajos de MapReduce, lo cual puede ser lento debido a la sobrecarga de MapReduce. En contraste, Impala utiliza su propio motor de ejecución para proporcionar consultas de baja latencia sin depender de MapReduce

Tema 8. Cloud computing

1. ¿Cómo se puede definir cloud computing? A. Es la interconexión de una serie de ordenadores. B. Es el proceso de planificar y ejecutar una serie de tareas C. Son una serie de servicios de computación ofrecidos a través de Internet. -> **Correcto** D. Ninguna de las respuestas anteriores son correctas.

2. ¿Qué modelos de servicio cloud existen? A. Público, privado e híbrido. -> **Correcto** B. IaaS, PaaS y SaaS. C. Microsoft Azure, Google Cloud Platform y Amazon Web Services. D. Servidores de cómputo, almacenamiento y bases de datos.

3. ¿Cuál de las siguientes propiedades no es una ventaja de cloud computing? A. Coste menor de infraestructura por economías de escala. B. Control total de la infraestructura que soporta los servicios. -> **Correcto** C. Flexibilidad a la hora de escalar la infraestructura necesaria. D. Alta disponibilidad de los servicios gracias a la replicación.

4. ¿Cuál de las siguientes opciones no es un tipo de nube? A. Nube pública. B. Nube privada. C. Nube secundaria. -> **Correcto** D. Nube híbrida

5. ¿Qué tipo de servicio no es habitual entre los servicios en la nube? A. Máquina virtual B. Máquina física. -> **Correcto** C. Almacenamiento virtual. D. Interconexión de servicios.

6. ¿Qué dos conceptos hacen posible los servicios de computación en la nube? A. Virtualización y disminución de costes. B. Disminución de costes y abstracción. C. Disminución de costes y flexibilidad. D. Abstracción y virtualización. -> **Correcto**

7. ¿Qué tarea reemplaza el uso de servicios en la nube? A. Compra e instalación de servidores. B. Actualización y mantenimiento de servidores. C. Dimensionamiento previo y adquisición de servidores para aumentar la capacidad según los requisitos de las aplicaciones. D. Todas las anteriores. -> **Correcto**

8. Relaciona los servicios de Microsoft Azure con su temática correspondiente es una imagen

9. Para ejecutar un clúster Hadoop en Microsoft Azure: A. Solo se puede usar el servicio HDInsight. B. Es obligatorio contratar una o varias instancias VM e instalar el clúster en ellas. C. Microsoft Azure no permite ejecutar clústeres Hadoop. D. Se puede usar el servicio HDInsight u optar por una alternativa IaaS. -> **Correcto**

10. Si se quieren utilizar servicios relacionados con machine learning en Microsoft Azure: A. Es necesario disponer de un equipo de expertos en machine learning que entiendan y puedan usar los servicios que provee Microsoft Azure. B. Microsoft Azure no proporciona ningún servicio de machine learning. Es necesario contratar un servicio de cómputo sobre el que instalar todo el entorno necesario para desarrollar modelos. C. Existen tanto opciones para conocedores de machine learning, que disponen de mayor flexibilidad para construir sus modelos, como servicios de inteligencia artificial que no requieren conocimientos de machine learning. -> **Correcto** D. Microsoft Azure no está diseñado ni orientado a ofrecer servicios de machine learning de ninguna forma

Videoclase 1. Cloud Computing : Introducción

- ¿Qué es el cloud computing según el contexto proporcionado?
- Un modelo de entrega de servicios tecnológicos a través de Internet que abstrae los detalles de implementación. => El documento define el cloud computing como un modelo que ofrece servicios tecnológicos a través de Internet, ocultando los detalles de implementación del sistema que proporciona el servicio, lo que permite a los usuarios acceder a recursos como almacenamiento, bases de datos y computación sin conocer las especificaciones de los sistemas físicos subyacentes.
- ¿Cuál de las siguientes opciones no es una ventaja del cloud computing mencionada en el documento?
- Necesidad de una gran inversión inicial en hardware. => El documento resalta que el cloud computing elimina la necesidad de una gran inversión inicial en hardware, ya que los servicios se pagan según el consumo y no requieren la compra de infraestructura física.
- ¿Qué permite la virtualización en el cloud computing?
- La compartición de recursos entre múltiples usuarios. => La virtualización en el cloud computing se refiere a la compartición de recursos, como computación y almacenamiento, entre múltiples usuarios, lo que permite escalar los servicios de manera ágil y según el consumo.

- ¿Qué característica del cloud computing mejora la recuperación ante desastres?
- La habilidad de detectar problemas y recuperarse rápidamente. => El documento menciona que la computación en la nube facilita la recuperación ante desastres y la continuidad del servicio de forma más sencilla y menos costosa, gracias a la capacidad de detectar problemas y recuperarse rápidamente.
- ¿Qué implica la escalabilidad en el contexto del cloud computing?
- La capacidad de un sistema para crecer y responder a un incremento de la demanda. => La escalabilidad en el cloud computing se refiere a la capacidad de un sistema para aumentar o disminuir sus recursos para responder a las variaciones en la demanda de peticiones, lo que se facilita gracias a la virtualización y la abstracción de los recursos.

Videoclase 2. Cloud Computing : Tipos

- ¿Cuál de los siguientes es un proveedor de nube pública?
- Google Cloud Platform (GCP).=>Google Cloud Platform (GCP) es uno de los principales proveedores de servicios de nube pública, junto con Amazon Web Services (AWS) y Microsoft Azure. Estos proveedores ofrecen recursos de computación, almacenamiento y red a través de Internet, accesibles para cualquier usuario mediante un modelo de pago por uso
- ¿Qué modelo de despliegue de nube combina nubes públicas y privadas?
- Nube híbrida. => La nube híbrida combina una nube pública con una privada mediante tecnologías que permiten compartir datos y aplicaciones entre ellas. Esto proporciona mayor flexibilidad, opciones de despliegue y ayuda a optimizar la infraestructura existente.
- ¿Cuál de los siguientes es un ejemplo de Software como Servicio (SaaS)?
- Dropbox => Dropbox es un ejemplo de Software como Servicio (SaaS), ya que proporciona almacenamiento y sincronización de archivos a través de Internet, gestionado y mantenido por el proveedor. Los usuarios acceden a las aplicaciones a través de un navegador sin necesidad de gestionar la infraestructura subyacente.
- ¿Qué modelo de servicio en la nube proporciona un entorno para desarrollar, probar y desplegar aplicaciones sin gestionar la infraestructura subyacente?

- PaaS => La Plataforma como Servicio (PaaS) proporciona un entorno completo bajo demanda para el desarrollo, pruebas, despliegue y gestión de aplicaciones sin que los desarrolladores tengan que preocuparse por la gestión de la infraestructura subyacente. Esto facilita la creación de aplicaciones web o móviles.

- ¿Cuál es la principal ventaja de utilizar una nube pública?

- Escalabilidad y flexibilidad. => La nube pública ofrece una gran escalabilidad y flexibilidad, permitiendo a las empresas ajustar rápidamente sus recursos según las necesidades sin una inversión inicial alta en infraestructura. Los recursos pueden ser accedidos desde cualquier lugar con conexión a Internet, y el mantenimiento es responsabilidad del proveedor

Videoclase 3. Cloud Computing : Microsoft Azure

- ¿Qué es Microsoft Azure?

- Una plataforma de computación en la nube. => Microsoft Azure es una plataforma de computación en la nube que ofrece un conjunto de servicios en continuo crecimiento para resolver diferentes problemas de negocio, desde servicios para albergar aplicaciones o páginas web hasta virtualización de servidores y servicios de almacenamiento.

- ¿Cuál de los siguientes servicios es parte de Azure Compute?

- Azure Virtual Machines.=>Azure Compute incluye servicios de cómputo bajo demanda como Azure Virtual Machines, que ofrecen máquinas virtuales Windows o Linux para ejecutar aplicaciones en la nube.

- ¿Qué servicio de Azure permite la gestión centralizada de cuentas de usuario y autenticar usuarios?

- Azure Active Directory. => Azure Active Directory es un servicio de Azure que proporciona gestión de identidad y acceso, permitiendo autenticar usuarios, gestionar roles y grupos de acceso, y configurar permisos de acceso para distintos servicios.

- ¿Cuál es la principal ventaja de usar Azure Virtual Machine Scale Sets?

- Permite escalar la capacidad de las máquinas virtuales según la demanda. => Azure Virtual Machine Scale Sets permiten escalar automáticamente la capacidad de las máquinas virtuales para manejar aumentos en la carga computacional, ajustándose a la demanda.

- ¿Qué tipo de almacenamiento proporciona Azure Blob Storage?
- Almacenamiento de objetos grandes como archivos de texto, vídeo o imagen. => Azure Blob Storage está orientado al almacenamiento de grandes objetos (blobs) como archivos de texto, vídeo o imagen, y datos no estructurados, siendo la opción más económica entre los servicios de almacenamiento de Azure.

Tema 9. Cloud computing II

1. ¿Cómo se distribuyen los recursos de la infraestructura de AWS? A. Se dividen en zonas, que, a su vez, tienen dos o más subzonas. B. Se dividen en regiones, que, a su vez, engloban dos o más zonas de disponibilidad. -> **Correcto** C. Se dividen en zonas de disponibilidad, que, a su vez, contienen dos o más centros de datos. D. No existe ninguna división, todos los recursos son globales e indistinguibles.

2. Una empresa quiere utilizar los servicios de AWS para almacenar datos personales y sensibles de sus clientes. ¿Cuál es el elemento más limitante a la hora de determinar dónde almacenar dichos datos? A. El coste, ya que, dependiendo de dónde se almacenen dichos datos, este puede ser mayor o menor. B. La latencia, ya que se tardaría mucho en obtener los datos si están almacenados lejos de donde se realiza la consulta. C. La legislación, porque, al ser información sensible, solo se pueden almacenar en lugares muy concretos para no incurrir en delitos. -> **Correcto** D. Todos los elementos anteriores tienen la misma importancia y hay que tenerlos en cuenta por igual

3. Una de las mayores ventajas de usar AWS como proveedor de servicios cloud es: A. Que tiene responsabilidad sobre toda la infraestructura, los servicios y los datos necesarios para desplegar nuestras aplicaciones. B. Que tiene responsabilidad sobre toda la seguridad concerniente a nuestra aplicación, desde el firewall al control de acceso. C. Que tiene la responsabilidad sobre la infraestructura y garantiza ciertos niveles de servicio al respecto. -> **Correcto** D. Todas las afirmaciones previas son correctas.

4. ¿Qué afirmación sobre las instancias de cómputo EC2 es incorrecta? A. Permiten elegir la imagen (AMI) que instalar en ellas de entre una colección predefinida o una proporcionada por el usuario. B. Poseen un conjunto predeterminado de configuraciones de cómputo, memoria y red, de donde escoger obligatoriamente la configuración predefinida que se desee. -> **Correcto** C. Se pueden contratar tantas instancias EC2 como se desee. D. Cada instancia EC2 está ligada a un servicio EBS para almacenamiento persistente.

5. ¿Cómo se interconectan los servicios AWS que contrata un usuario? A. Mediante una red global que comparten todos los servicios contratados por todos los usuarios en AWS. B. Mediante una red propia del usuario que conecta las direcciones IP de los servicios contratados. C. Los distintos servicios contratados son independientes y autocontenidos, por lo que no necesitan ni pueden comunicarse con otros. D. Mediante el servicio de interconexión AWS VPC. -> **Correcto**

6. Indica qué caso de uso no es propio de S3: A. Sistema de arranque de una instancia EC2. -> **Correcto** B. Almacenamiento de ficheros accesibles desde instancias EC2. C. Almacenamiento de ficheros accesibles desde un navegador web. D. Almacenamiento de archivo de ficheros de escaso acceso.

7. Se quiere desplegar una base de datos relacional de forma rápida y que no suponga una carga de mantenimiento para el departamento de IT, más allá de la gestión de los datos contenidos. ¿Qué servicio AWS escogerías? A. Instancia EC2 e instalación de MySQL. B. AWS RDS. -> **Correcto** C. AWS DynamoDB. D. AWS ECS e instalación de MySQL.

8. ¿Cuál de las siguientes opciones es la mejor para desplegar un servicio de almacenamiento distribuido en AWS? A. Varias instancias EC2 sobre las que el usuario instala un clúster Hadoop, que incluye HDFS. B. Varios contenedores ECS sobre los que el usuario instala un clúster Hadoop, que incluye HDFS. C. Un clúster EMR, con su propio sistema de almacenamiento distribuido. -> **Correcto** D. Un clúster EMR, con el sistema de almacenamiento HDFS.

9. Amazon SageMaker es un servicio de AWS destinado a: A. Construir y entrenar modelos de machine learning desde cero. -> **Correcto** B. Utilizar modelos de machine learning preconstruidos. C. Realizar consultas interactivas sobre grandes conjuntos de datos. D. Catalogar todos los datos existentes en los diferentes servicios AWS

10. Si se quieren manejar flujos de datos en tiempo real, ¿qué servicio AWS no sería adecuado? A. Amazon Kinesis Streams. B. Amazon MSK. C. Amazon Redshift. -> **Correcto** D. Instancias EC2 con Kafka instalado

Videoclase 1.

- Qué es Amazon Web Services (AWS)?

- Una plataforma de computación en la nube. => WS es una plataforma líder en servicios de computación en la nube, proporcionando soluciones a empresas, startups, el sector público y particulares. Ofrece un amplio conjunto de servicios que incluyen almacenamiento, procesamiento, redes y seguridad, entre otros
- ¿Qué concepto describe mejor la organización geográfica de los servicios de AWS?
- Regiones y zonas de disponibilidad (AZ). => AWS organiza sus servicios en regiones geográficas, cada una de las cuales contiene varias zonas de disponibilidad (AZ). Cada región es una colección de centros de datos separados geográficamente, pero interconectados para proporcionar alta disponibilidad y resiliencia.
- ¿Cuál es una consideración importante al elegir una región de AWS para desplegar aplicaciones?
- La localización de los usuarios finales y la latencia
- ¿Qué modelo de despliegue permite AWS para flexibilizar la migración de servicios entre servidores propios y la nube?
- Despliegue híbrido =>
- ¿Cuál de los siguientes es un servicio IaaS proporcionado por AWS que puede utilizarse para desplegar soluciones de big data? => Amazon EC2 es un servicio de infraestructura como servicio (IaaS) que proporciona capacidad de cómputo escalable en la nube. Es ideal para desplegar soluciones de big data, ofreciendo flexibilidad y control sobre la infraestructura subyacente.
-

Videoclase 2. Servicios

- ¿Cuál de los siguientes servicios de AWS se utiliza para la encriptación y gestión de claves?
- AWS Key Management Service (KMS).
- ¿Qué servicio de AWS permite a los usuarios ejecutar código en respuesta a eventos sin necesidad de gestionar servidores?
- AWS Lambda
- ¿Cuál de las siguientes clases de almacenamiento de Amazon S3 está diseñada para datos a los que no se accede de forma frecuente y se almacenan en una única AZ?

- Amazon S3 One Zone-Infrequent Access (IA) => está diseñada para datos que no se acceden frecuentemente y se almacenan en una única zona de disponibilidad (AZ), ofreciendo un costo más bajo con la condición de que los datos pueden perderse en caso de un fallo en el centro de datos.

- ¿Cuál de los siguientes servicios de AWS proporciona monitoreo continuo y detección de amenazas?

- Amazon GuardDuty => es un servicio de detección inteligente de amenazas que ofrece monitoreo continuo para identificar comportamientos sospechosos y proteger las aplicaciones alojadas en AWS.

- ¿Cuál de los siguientes servicios de almacenamiento en AWS es similar a un NAS y permite el acceso simultáneo a múltiples instancias EC2?

- Amazon EFS. => Amazon Elastic File System (EFS) => proporciona almacenamiento para archivos en la nube con acceso masivo y paralelo compartido, similar a un Network Attached Storage (NAS), permitiendo que múltiples instancias EC2 accedan al mismo sistema de ficheros simultáneamente.

Videoclase 3. Otros servicios

- ¿Qué servicio de AWS facilita la ejecución de aplicaciones de Big Data en Hadoop sin que el usuario tenga que gestionar las instancias EC2?

- Amazon EMR. => Amazon EMR (Elastic MapReduce) es una plataforma que proporciona tecnologías Hadoop sobre instancias EC2 gestionadas por AWS, permitiendo al usuario centrarse en el uso de los servicios sin tener que gestionar las instancias ni crear el clúster por sí mismo

- ¿Cuál de los siguientes servicios de AWS es específico para la extracción, transformación y carga (ETL) de datos?

- AWS Glue. => AWS Glue es un servicio serverless enfocado en tareas ETL (extract, transform, load), proporcionando un editor visual y un catálogo de datos para gestionar y transformar datos almacenados en diversas fuentes y formatos.

- ¿Cuál de los siguientes servicios de AWS es específico para la extracción, transformación y carga (ETL) de datos?

- Realizar transformaciones de datos casi en tiempo real. => Kinesis Data Firehouse permite gestionar datos casi en tiempo real, con capacidades de transformación de datos y escalado automático, ideal para cargar flujos de datos en servicios como S3, Redshift y ElasticSearch

- ¿Qué característica diferencia a Amazon Redshift de Amazon RDS?

- Redshift es un sistema OLAP mientras que RDS es OLTP. => Amazon Redshift es un sistema de data warehousing orientado a columna optimizado para consultas OLAP (OnLine Analytical Processing), mientras que Amazon RDS está orientado a OLTP (OnLine Transaction Processing), permitiendo consultas transaccionales.

- ¿Cuál es el principal caso de uso de Amazon Athena?

- Consultas serverless a datos almacenados en S3. => Amazon Athena es un servicio serverless que permite realizar consultas interactivas a datos almacenados en S3 sin necesidad de cargarlos, soportando muchos formatos y facilitando el análisis interactivo de datos.

Tema 10. Cloud computing III

1. ¿Cómo se organizan los recursos, los servicios y las políticas de seguridad que contrata y configura un usuario u organización en Google Cloud? A. En folders, que contienen proyectos. -> **Correcto** B. En proyectos, que contienen folders. C. En proyectos y zonas. D. En zonas, que contienen diferentes folders.

2. Si queremos aumentar la disponibilidad de un servicio GCP, ¿qué debemos hacer? A. Desplegarlo en la región más cercana a su uso. B. Desplegarlo en una región que no presente problemas legales con la información que gestiona. C. Desplegarlo como recurso regional o multirregional. -> **Correcto** D. Desplegar una instancia VM que esté siempre ejecutándose.

3. Elige la respuesta incorrecta: A. GCP proporciona una serie de servicios de AI bajo AI Platform, para usuarios no expertos en el dominio, los cuales quieran usar AI en sus aplicaciones sin desarrollar ningún modelo. B. GCP proporciona una serie de servicios de AI bajo AI Platform, para usuarios expertos en el dominio que quieran usar AI en sus aplicaciones desarrollando sus propios modelos. C. GCP proporciona una serie de servicios de AI bajo Cloud AI Building Blocks, para usuarios no expertos en el dominio, los cuales quieran usar AI en sus aplicaciones sin desarrollar ningún modelo. -> **Correcto** D. Entre los servicios AI para uso directo, se pueden encontrar herramientas de clasificación de imágenes o vídeo, o traductores entre diferentes idiomas

4. En cuanto a la seguridad, ¿qué esquema sigue GCP? A. Un esquema de seguridad compartida, donde GCP se hace siempre cargo de todos los niveles, excepto de los datos. B. Un esquema de seguridad compartida, donde GCP se hace cargo de ciertos niveles, que dependen del servicio desplegado. -> **Correcto** C. El usuario debe hacerse cargo de la seguridad de todo el sistema, que sigue un modelo de cuatro capas. D. Un esquema de seguridad compartida de cuatro capas, donde el usuario solo se hace cargo de la capa a más alto nivel y Google Cloud, de todos los aspectos de las otras tres.

5. ¿Qué opción es más interesante para ejecutar tareas cortas y no críticas, que se podrían repetir si fuera necesario? A. Instancias VM normales. B. Preemptible instances VM. -> **Correcto** C. Sole-tenant VM instances. D. One use VM instances.

6. Una empresa quiere almacenar los datos históricos de las nóminas de los empleados, con el único objetivo de hacer frente a una posible auditoría en los cinco años siguientes al pago de cada nómina ¿Qué opción de almacenamiento de GCP es la más adecuada en cuanto a acceso y coste? A. Cloud Storage Coldline. B. BigTable. C. Cloud Storage Archive. -> **Correcto** D. Cloud Persistent Disks

7. Cuando se crea un proyecto en Google Cloud, ¿cómo se interconectan los servicios que engloba? A. Se crea una VPC, que contiene el rango de direcciones IP que se asignan los servicios. B. Hay que definir siempre manualmente las subredes de una VPC para tener disponibles direcciones IP que asignar a los servicios. C. Se crea automáticamente una subred dentro de la VPC, que contiene el rango de direcciones IP disponibles para asignar a los servicios. -> **Correcto** D. Un proyecto solo se puede interconectar con otro, pero los servicios dentro de un proyecto no se interconectan y, por tanto, no se necesitan direcciones IP.

8. Elige la respuesta incorrecta: Si quisiéramos desplegar un clúster Hadoop en GCP, podríamos… A. Usar varias instancias VM configuradas manualmente como clúster e instalar las herramientas del ecosistema Hadoop deseadas. B. Usar el servicio Dataproc. C. Usar el servicio Dataflow. -> **Correcto** D. Usar varios contenedores GKE configurados manualmente como clúster e instalar las herramientas del ecosistema Hadoop deseadas.

9. ¿A qué base de datos de código libre se asemeja BigTable? A. Es un motor propietario único de Google, muy diferente a cualquier otra base de datos existente. B. MongoDB. C. Cassandra. D. HBase. -> **Correcto**

Nota: BigTable se asemeja a la base de datos de código libre HBase. Ambas son bases de datos NoSQL de tipo columnar que están diseñadas para manejar grandes volúmenes de datos. HBase, de hecho, se inspiró directamente en el modelo de datos de BigTable, que fue descrito en un paper de Google en 2006.

10. Relaciona cada servicio GCP con el que sería su equivalente en proyectos Apache

imagen se debe poner

Videoclase 1. Cloud Computing : Introducción

- ¿Qué característica de Google Cloud Platform (GCP) permite escalar recursos según la demanda?
- Escalabilidad => Escalabilidad. La escalabilidad es la capacidad de aumentar o disminuir los recursos de acuerdo con la demanda, una característica clave de GCP.
- ¿Cuál es la estructura que organiza los servicios de Google Cloud Platform para ofrecer flexibilidad, disponibilidad y rendimiento?
- Todas las anteriores.
- ¿Qué beneficio ofrece la implementación de servicios en múltiples zonas dentro de una región?
- Alta disponibilidad. => Alta disponibilidad. Desplegar servicios en múltiples zonas dentro de una región proporciona redundancia, lo que asegura una mayor disponibilidad y resistencia ante fallos.
- ¿Cuál de los siguientes es un ejemplo de región de Google Cloud Platform?
- asia-east1. => asia-east1. asia-east1 (Taiwán) es un ejemplo de región mencionada en GCP, proporcionando servicios cerca de los usuarios para reducir la latencia y cumplir con regulaciones locales de datos.
- ¿Cuál es el propósito principal de las multirregiones en Google Cloud Platform?
- Ofrecer durabilidad y disponibilidad de datos más alta. Las multirregiones están diseñadas para ofrecer una durabilidad y disponibilidad de datos superior, adecuada para el almacenamiento y replicación de datos críticos.

Videoclase 2. Servicios

- ¿Qué servicio de Google Cloud Platform permite la orquestación de contenedores con Kubernetes?
- Google Kubernetes Engine. => Google Kubernetes Engine (GKE) es el servicio específico de Google Cloud Platform diseñado para la orquestación de contenedores utilizando Kubernetes, permitiendo una gestión simplificada y escalabilidad para implementaciones de microservicios.
- ¿Cuál es una característica destacada de Cloud Functions en GCP?
- Ejecución basada en eventos. => Cloud Functions es un servicio sin servidor de Google Cloud Platform que permite la ejecución de funciones en respuesta a eventos específicos, facilitando la automatización de tareas y el procesamiento en tiempo real.
- ¿Qué servicio de almacenamiento en GCP está diseñado para bases de datos relacionales gestionadas?
- Cloud SQL. => Cloud SQL es el servicio de Google Cloud Platform que ofrece bases de datos relacionales gestionadas, compatible con MySQL, PostgreSQL y SQL Server, y es ideal para aplicaciones empresariales que requieren una alta durabilidad y acceso global.
- ¿Qué característica distingue a la Virtual Private Cloud (VPC) de GCP?
- Aislamiento y control de red. => Aislamiento y control de red. La Virtual Private Cloud (VPC) de Google Cloud Platform permite la creación de redes privadas virtuales, ofreciendo características de aislamiento y control de red para implementaciones seguras y entornos híbridos.
- ¿Cuál es un caso de uso típico para Google Compute Engine en GCP?
- Aplicaciones que requieren recursos específicos. => Google Compute Engine proporciona máquinas virtuales escalables que son altamente personalizables y adecuadas para aplicaciones que requieren recursos específicos, como un alto rendimiento y personalización

Videoclase 3. Otros servicios

- ¿Cuál de los siguientes servicios de Google Cloud Platform (GCP) se utiliza para el procesamiento de datos en tiempo real y por lotes?

- Dataflow => Dataflow es un servicio basado en Apache Beam que permite el procesamiento de datos tanto en tiempo real como en lotes. Está diseñado para optimizar mediante grafo acíclico dirigido (DAG) y soporta la paralelización y el ordenamiento topológico.
- ¿Qué servicio de GCP proporciona un sistema de mensajería y transmisión de datos en tiempo real con alta disponibilidad y escalabilidad?
- Pub/Sub. => es el servicio de mensajería y transmisión de datos en tiempo real de GCP, conocido por su alta disponibilidad y escalabilidad. Es utilizado para la integración de aplicaciones y el streaming de eventos.
- ¿Cuál es una característica destacada de BigQuery en GCP?
- Interfaz SQL para análisis de datos a gran velocidad y escala.. => nterfaz SQL para análisis de datos a gran velocidad y escala. BigQuery es conocido por ser un sistema de data warehousing que utiliza una interfaz SQL para realizar análisis de datos de manera rápida y a gran escala, permitiendo la inferencia de esquemas y carga de datos en bloque o en tiempo real.
- ¿Cuál de los siguientes es un caso de uso de dataflow en GCP?
- Procesamiento de eventos y ETL. => Procesamiento de eventos y ETL. Dataflow se utiliza principalmente para el procesamiento de eventos y ETL (Extracción, Transformación y Carga) debido a su capacidad para manejar y procesar grandes volúmenes de datos en tiempo real y en batch.
- ¿Qué servicio de GCP está diseñado para la visualización de datos y la creación de informes, y se integra con varias fuentes de datos?
- Data Studio. => Data Studio. Data Studio es una herramienta de GCP que permite la visualización de datos y la creación de informes interactivos, con una interfaz intuitiva y la capacidad de integrarse con diversas fuentes de datos, facilitando la creación de dashboards.