***Business Intelligence***

***&***

***“Big” Data***

***TRABAJO FINAL – Entrega 06.03.2023***

|  |  |
| --- | --- |
| **Profesor** | ***David Díaz*** [***ddiaz@unegocios.cl***](mailto:ddiaz@unegocios.cl) |
| ***Alumnos*** | ***Aníbal Quiroz***  ***Rodrigo Valdés*** |

|  |
| --- |
| **Pregunta 1. Desarrollo**  Explique con sus palabras cuáles son las principales diferencias y similitudes entre:  Una base de datos transaccional  Un Data Warehouse  Un Data Lake  Un Lake House  Proponga un ejemplo y caso de uso para cada una de ellas |

| **Criterio** | **Base de datos transaccional** | **Data Warehouse** | **Data Lake** | **Lake House** |
| --- | --- | --- | --- | --- |
| ***Tipo de datos*** | *Contiene datos estructurados* | *Contiene datos estructurados* | *Contiene datos estructurados y no estructurados* | *Contiene datos estructurados y no estructurados* |
| ***Conveniencia*** | *Conveniente para operaciones de escritura (Write/Update)* | *Conveniente para operaciones de lectura (Read)* | *Conveniente para operaciones de lectura y escritura, pero los datos a leer están “Raw” o sin preprocesar, requiere mayor manejo.* | *Conveniente para operaciones de lectura y escritura, con lo mejor del preprocesamiento del DWH y la flexibilidad del Data Lake.* |
| ***Uso principal*** | *Transacciones*  *Ej. Supermercados* | *Business intelligence, analítica de datos o visualización* | *Machine learning, análisis predictivos, visualización de datos, BI con preprocesamiento* | *Business intelligence, analítica de datos, visualización,*  *Machine learning y análisis predictivo* |
| ***Procesamiento*** | *Existe en menor medida, por ejemplo virtual/stored pre-calculated columns, pero suele afectar la eficiencia* | *ETL (Extracción, transformación y carga), la información se pre-procesa y está lista para su visualización* | *ELT (Extracción, transformación y carga), la información se procesa y estructura solamente cuando esta se solicita* | *ETL (Extracción, transformación y carga)*  *La información se estructura en warehouse sobre los lake, lo que permite entregarle la versatilidad para poder ser trabajados en ML y en BI* |
| ***Usuario objetivo*** | *Básico (Usuario que maneje entrada de datos) Ej. Operador* | *Intermedio (Usuario que sepa interpretar los datos) Ej.Manager* | *Avanzado*  *(Usuario que sepa pre-procesar los datos) Ej. Data Scientist* | *Todas las anteriores* |
| ***Temporalidad*** | *Suelen ser almacenadas un par de años* | *Suelen contener los datos históricos* | *Depende de la complejidad de los datos que se están almacenando y la capacidad a nivel de recursos de la organización* | *Igual que el data lake, depende, pero en esta oportunidad, los warehouse pueden mantener la información histórica independientemente que los lake sean eliminados.* |
| ***Oportunidad de los datos*** | *Tiempo real* | *Depende del ETL y su configuración* | *Tiempo real* | *Mixto* |
| ***Ejemplo*** | *Compras en supermercados* | *Warehouse de Finanzas* | *Sistema de vigilancia* | *Facebook* |
| ***Caso de uso*** | *Anotar las transacciones de las compras de los clientes* | *Mantener una BD pre-procesada para la toma de decisiones* | *Anotar de forma no estructurada segmentos de grabaciones y de forma estructurada las detecciones de movimiento* | *Todas las anteriores* |

|  |
| --- |
| **Pregunta 2. Desarrollo**   1. **Explique con sus palabras cuales son los *drivers* o causas que hacen necesario la utilización de “clústers de computadoras”** 2. **Refiérase especialmente al uso de HPC (clústers intensivos en cómputo).s al uso de clústers para Big Data (tipo Hadoop o Spark).**   **Proponga un ejemplo o caso de uso para cada uno de ellos.** |

|  |
| --- |
| ***Respuesta A)***  *Considerando se podría entender a un clúster de computadoras como un grupo de dos o más computadoras que se ejecutan en paralelo para lograr un objetivo común. En este sentido un driver o causa relevante de es que la utilización de clúster puede permitir que las cargas de trabajo que consisten en una gran cantidad de tareas individuales se distribuyan entre la capacidad de las computadoras. Como resultado, estas tareas pueden aprovechar la memoria combinada y la potencia de procesamiento de cada computadora para aumentar el rendimiento general.*  *Uno entendería que un usuario que acceda al clúster no debería necesitar saber si el sistema es un clúster o una máquina individual, dado que los nodos individuales deben estar conectados en una red para permitir la comunicación entre ellos. El software de clúster de computadoras se puede usar para unir los nodos y formar un clúster.*  *Se entiende que los tipos de clúster han surgido para apoyar los procesos en cuanto a alta disponibilidad a través de tolerancia a fallas, equilibrio de carga y computación de alto rendimiento. Por ejemplo: una aplicación que se ejecuta en una sola máquina tiene un único punto de falla, lo que hace que la confiabilidad del sistema sea deficiente. Si la máquina que aloja la aplicación deja de funcionar, casi siempre habrá tiempo de inactividad mientras la infraestructura se recupera.* |

|  |
| --- |
| ***Respuesta B.1)***  *HPC es una tecnología que utiliza grupos de potentes procesadores, trabajando en paralelo, para procesar conjuntos de datos masivos multidimensionales (big data) y resolver problemas complejos a velocidades extremadamente altas.*  *Las cargas de trabajo de HPC descubren nuevos conocimientos importantes que hacen avanzar el conocimiento humano y crean una ventaja competitiva significativa. Por ejemplo, HPC se utiliza para secuenciar el ADN, automatizar el comercio de acciones y ejecutar algoritmos y simulaciones de inteligencia artificial (IA), como los que habilitan los automóviles autónomos, que analizan terabytes de transmisión de datos desde sensores IoT, radares y sistemas GPS en tiempo real. para tomar decisiones en fracciones de segundo.*  *Un sistema informático estándar resuelve problemas principalmente mediante la informática en serie: divide la carga de trabajo en una secuencia de tareas y luego ejecuta las tareas una tras otra en el mismo procesador. Por el contrario, HPC:1) Ejecuta múltiples tareas simultáneamente en múltiples servidores o procesadores de computadoras, 2) Consta de varios servidores informáticos de alta velocidad conectados en red, con un programador centralizado que gestiona la carga de trabajo informática paralela y 3) Los recursos informáticos de un clúster HPC (redes, memoria, almacenamiento y sistemas de archivos) son componentes de alta velocidad, alto rendimiento y baja latencia.*  *Un ejemplo de uso de HPC es el pronóstico del tiempo y el modelado climático, los cuales implican el procesamiento de grandes cantidades de datos meteorológicos históricos y millones de cambios diarios en puntos de datos relacionados con el clima. Se puede agregar también para el caso de el procesamiento de datos sísmicos.* |

|  |
| --- |
| ***Respuesta B.2)***  *Por otro lado Hadoop aporta a la computación clásica de alto rendimiento (HPC) es la localidad de datos. Hadoop lleva la computación a los datos. (HPC compensa al tener interconexiones más rápidas como almacenamiento de gran ancho de banda).*  *El gran avance que Spark trajo a Hadoop es almacenar datos en la memoria RAM de cada nodo en lugar de en el disco de cada nodo. El aprovechamiento de la localidad de los datos por parte de Spark es muy similar al de Hadoop: es decir, el cálculo se asigna para que ocurra donde residen los datos.*  *Un ejemplo de uso de HADOOP o SPARK*  *Casos de uso de Hadoop - Hadoop es más efectivo para escenarios que involucran lo siguiente:*   * *Procesamiento de grandes conjuntos de datos en entornos donde el tamaño de los datos supera la memoria disponible* * *Procesamiento por lotes con tareas que aprovechan las operaciones de lectura y escritura del disco* * *Construir una infraestructura de análisis de datos con un presupuesto limitado* * *Completar trabajos que no son urgentes* * *Análisis de datos históricos y de archivo*   *Casos de uso de chispas - Spark es más efectivo para escenarios que involucran lo siguiente:*   * *Tratar con cadenas de operaciones paralelas mediante el uso de algoritmos iterativos* * *Lograr resultados rápidos con cálculos en memoria* * *Análisis de datos de flujo en tiempo real* * *Procesamiento gráfico paralelo para modelar datos* * *Todas las aplicaciones de aprendizaje automático* |

|  |
| --- |
| **Pregunta 3. Ejercicio EDA con Spark SQL**  **Utilizando Apache Spark (en Colab, o DataBricks, o local, …), las librerías de SparkSQL y el dataset SII\_roles\_avaluo\_table.zipLinks to an external site.…**  **Escriba 3 nuevas consultas (queries) a la data:**   * **Debe usar a lo menos un Left Join** * **Debe usar a lo menos un GroupBy** * **Debe usar un a lo menos Where con dos condiciones** * **Debe ordenar sus resultados de manera descendiente** |

|  |  |
| --- | --- |
| Se adjunta Archivo (códigos (link a colab o notebook .ipynb)) | https://colab.research.google.com/drive/1vUN\_5NKzEMo6V2SVj1nahHwKrZXiq\_qj?usp=sharing |

|  |
| --- |
| **Pregunta 4. Opción A. Ejercicio Clustering Bancario**  Utilizando Apache Spark (en Colab, o DataBricks, o local, …), las librerías de SparkML y el dataset créditos\_bancarios.xlsx …  Realice una segmentación (usando K-means) de la cartera de clientes que incluya a lo menos 5 variables de su interés.  En sus resultados comente respecto de:  A) qué tipo de preprocesamientos fue necesario realizarle a los datos, o sino fue necesario, el por qué.  B) Cómo se determinó el número óptimo de clusters a utilizar  C) La estadística descriptiva de los segmentos encontrados y qué nombre “comercial” le pondría al segmento dadas dichas características  D) Acciones de negocios que podrían ser relevantes a sugerir para los segmentos encontrados |

|  |  |
| --- | --- |
| Se adjunta Archivo (códigos (link a colab o notebook .ipynb)) | https://colab.research.google.com/drive/1CZkHZIRQy62jDKpN43RovpTC-5zTfLLz?usp=sharing |
| Respuestas | *A) No fue necesario realizarle preprocesamientos a los datos debido a que ya los datos estaban en formato numérico o dígitos, por lo que solo bastaba elegir los campos considerados relevantes. Solo por tratarse de spark, se creo una columna “features” (que concatena 5 campos), y se incluye la instrucción de transformar y mostrar.*  *B) Se determinó el número óptimo de clusters a utilizando K-Means. La silueta se calcula para saber si el clustering fue bueno. Se obtuvo “Silhouette with squared euclidean distance = 0.7084788086509726”, cercano a 1 significa que son mas cercanos los valores.*  *C) De todas maneras se logra tener 10 clusters / segmentos encontrados, que se concentra principalmente los segmentos con predicción 0, 5, 6 y 8,*  D) *Las acciones de negocios que podrían ser relevantes a sugerir para los 10 donde el rango de edad está entre los 34 y 35 años, donde el tiempo de duración en los trabajos es entre 1 y 4 años , con ahorros ... < 100 DM y con building soc. savings agr./life insurance .* |

|  |
| --- |
| **Pregunta 4. Opción B. Ejercicio Modelamiento Supervisado clasificación**  **Utilizando Apache Spark (en Colab, o DataBricks, o local, …), las librerías de SparkML y el dataset créditos\_bancarios.xlsx …**  **Cree un modelo de predicción de la variable “credit\_risk”. Pruebe con al menos 3 algoritmos supervisados distintos, por ej, regresión logística, árboles de decisión y random forests:**  **En sus resultados comente respecto de:**  **A) qué tipo de preprocesamientos fue necesario realizarle a los datos, especialmente respecto de si fue necesario balancear la data**  **B) Los resultados de su modelo tanto en muestra de training como de testing.**  **C)Cree una tabla donde se puedan comparar las métricas de recall, precisión y accuracy para ambas clases “good” y “bad” payer (pagador). Comente y justifique cual de los 3 algoritmos funcionó mejor y por qué.**  **D) Realice una breve discusión de cómo se podría implementar dicho modelo en la práctica.** |

Solo se realiza **4opciónA** no **4opciónB**

**Condiciones de entrega**

Puede ser realizado de manera **individual o en parejas**

Debe entregar todos sus códigos (link a colab o notebook .ipynb)

Los códigos deben venir explicados y documentados

Puede hacer los supuestos que estime convenientes, pero éstos deben estar documentados

Se debe entregar vía Canvas del curso hasta antes de la fecha que sea estipulada por el Profesor.