******

*Trabajo Final:*

*“Datos masivos y minería de datos”*

Profesor:

- David Diaz

Integrantes:

- Cristian Hinojosa

- Pablo Chaparro

Marzo 2023

**Pregunta 1. Desarrollo**

Explique con sus palabras cuáles son las principales diferencias y similitudes entre:

Una base de datos transaccional

Una base de datos (BD) transaccional guarda datos estructurados, organizados en tablas, filas y columnas, el principal objetivo es asegurar la transacción dentro de una base relacional.

Un ejemplo típico de este tipo de BDs es el registro de ventas de una empresa donde cada fila corresponde a una venta y que guarda todos los datos de la transacción en columnas distintas

Un Data Warehouse

Un Data Warehouse es una BD orientada a la lectura, tiene una estructura OLAP, es capaz de realizar procesamiento de datos y entrega información útil para el análisis.

Un ejemplo común de data warehouse es la generación de campañas de venta donde a través del procesamiento y análisis de la información se perfilan ciertos productos para un segmento determinado de clientes.

Un Data Lake

Un Data Lake es un tipo de BD que tiene la capacidad de guardar (solo guardar) datos estructurados y no estructurados en grandes cantidades, de forma distribuida y a escala, sin aplicar algún tipo de transformación ni gobernanza, además, no aseguran las propiedades ACID sobre la información.

Un ejemplo claro de un Data Lake puede ser una red social que debe almacenar fotos, videos, comentarios (no editables), audios, etc…

Un Lake House

Un Lake House es un sistema que combina las características de un Data Lake con los Data Warehouse, garantizando todas las propiedades ACID + gobernanza de los datos, este tipo de BDs también permite actualizar y eliminar datos.

Los ejemplos más interesantes sobre el uso de Lake House están en el ámbito de la IA y ML.

Tabla resumen sobre la comparación de las bases de datos:

| **Característica** | **BD transaccional** | **Data Ware House** | **Data Lake** | **Lake House** |
| --- | --- | --- | --- | --- |
| Almacenar grandes volúmenes de datos |  | x | x | x |
| Guardar datos estructurados | x | x | x | x |
| Guardar datos NO estructurados |  |  | x | x |
| Actualizar Datos | x | x |  | x |
| Eliminar Datos | x | x |  | x |
| Transformación de datos |  | x |  | x |
| Utilización de datos en tiempo real |  |  | x | x |

**Pregunta 2. Desarrollo**

Explique con sus palabras cuales son los *drivers* o causas que hacen necesario la utilización de “clústers de computadoras”

Uno de los principales *drivers*  que hacen necesario la utilización de *Cluster* de computadoras es la necesidad de resolver tareas que requieren mucha memoria, mucho poder de cómputo o procesar grandes cantidades de datos, los computadores normales por lo general tienen capacidades limitadas de cómputo y eso se traduce en procesamiento secuencial generalmente. Por otro lado, al tener muchas computadoras conectadas entre sí y gestionadas por algún tipo de software estas pueden ejecutar tareas masivas de forma distribuida o de forma paralela.

Refiérase especialmente al uso de HPC (clústers intensivos en cómputo) vs al uso de clústers para Big Data (tipo Hadoop o Spark)

Los HPC son por lo general grandes computadores con mucha capacidad de cómputo, miles de CPU y miles de GPU trabajando en equipo, siendo capaces de procesar millones de operaciones (Tflop). Los HPC están orientados a ejecutar tareas que requieren mucho cómputo con poca cantidad de datos. En cambio, Big Data, en oportunidades no requiere mucho cómputo, sino más bien, necesita procesar grandes cantidades de datos y es aquí donde se producen las principales diferencias ya que el componente tecnológico de ambas BDs llamado *JOB Schedulers* es distinto dependiendo que se quiera ejecutar, por lo general HPC utiliza Slurm y Big Data (Hadoop para este ejemplo) utiliza Yarn, ambas son opensource y corren sobre Linux, HPC soporta todos los lenguajes de programación y Yarn solo JAVA y Python.

En resumen, HPC cuenta con más características que Big Data pero tienen orientaciones muy distintas, el primero hacia el cómputo y el segundo hacia los datos.

Proponga un ejemplo o caso de uso para cada uno de ellos.

Un ejemplo de HPC es la simulación del curso de una tormenta incluyendo los posibles daños de esta, donde a partir de pocos datos el HPC puede generar miles de simulaciones y por consiguiente como resultado tener mucha información, en este caso se necesita mucho cómputo para generar todas las simulaciones.

En cambio en el caso de Big Data un ejemplo podría ser el análisis de calidad de una fruta a partir de fotografías tomadas por drones a cada fruta en un campo de muchas hectáreas. El proceso de análisis de cada foto es simple, por lo que el consumo de cómputo es bajo sin embargo por la gran cantidad de datos a procesar (millones de frutos) es necesario procesar con tecnología orientada al Big Data

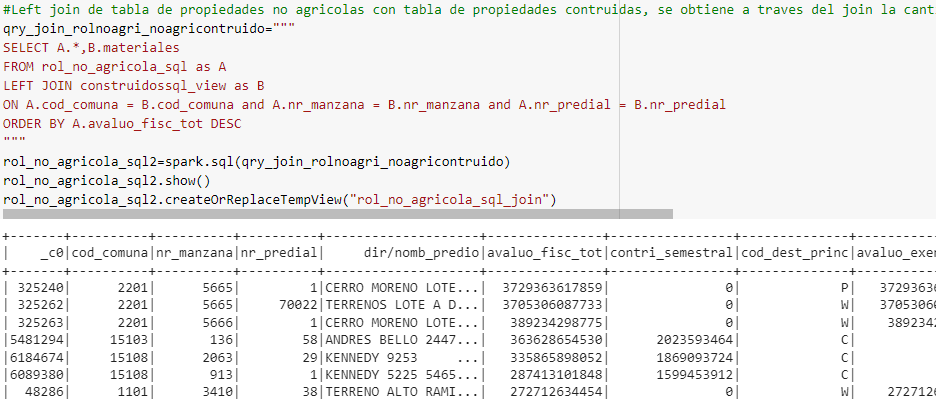
**Pregunta 3. Ejercicio EDA con Spark SQL**

Utilizando Apache Spark (en Colab, o DataBricks, o local, …), las librerías de SparkSQL y el dataset [*SII\_roles\_avaluo\_table.zip Links to an external site.*](https://www.dropbox.com/s/ianxkffu14dp0ru/SII_roles_avaluo_table.zip)…

Escriba 3 nuevas consultas (queries) a la data:

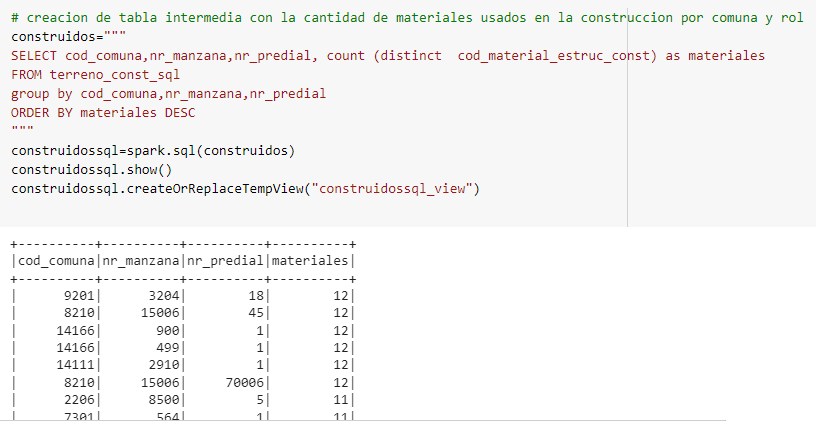
* Debe usar a lo menos un Left Join

Left join de tabla de propiedades no agrícolas con tabla de propiedades construidas, se obtiene a través del join la cantidad de materiales diferentes utilizados.



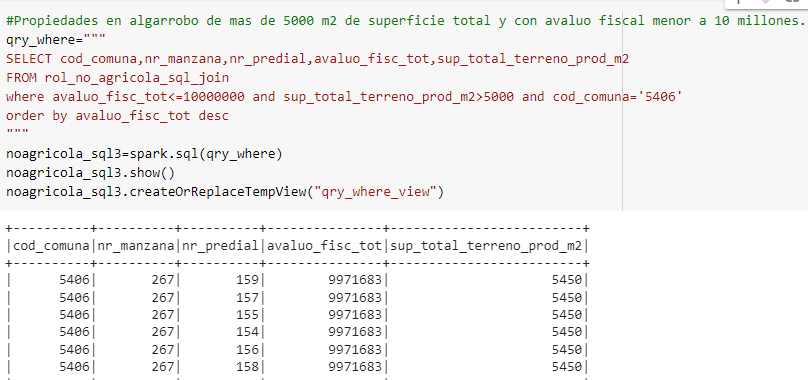
* Debe usar a lo menos un GroupBy

Creación de tabla intermedia con la cantidad de materiales usados en la construcción por comuna y rol



* Debe usar un a lo menos Where con dos condiciones

Propiedades en algarrobo de más de 5000 m2 de superficie total y con avalúo fiscal menor a 10 millones.



* Debe ordenar sus resultados de manera descendente

Todas las consultas quedaron ordenadas de manera descendente.

**Pregunta 4. Opción A. Ejercicio Clustering Bancario (K-Means)**

El ejercicio fue desarrollado en el ambiente Colab utilizando librerías de SparkML y el dataset *créditos\_bancarios.xlsx.*

A) qué tipo de preprocesamiento fue necesario realizarle a los datos, o si no fue necesario, el por qué.

* **Carga de datos**: lo primero fue cargar el archivo .xlsx dentro de un archivo .rar y subirlo a Dropbox, de esta manera se logra trabajar de manera más automatizada sin necesidad de estar cargando el archivo cada vez que se corre el programa a través de google drive u otro sistema que requiera validación.
* **Análisis exploratorio**: De la base de datos “SouthGermanCredit” se identifican los tipos de variables que existen, considerando de antemano que solo se pueden elegir 5 variables:

Como se observa en la siguiente tabla sólo 3 variables son numéricas (amount, age, duration) y el resto es categórica de acuerdo a la descripción que incluye el archivo excel.

Los tipos de variables son muy relevantes para realizar el clustering usando k-means dado que para que tenga sentido hacerlo todas las variables deben ser numéricas y para los casos de variables categóricas se deben transformar a dummies.

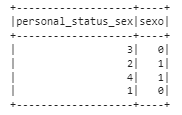
Por lo anterior se determina utilizar las 3 variables numéricas y elegir 2 adicionales en función de la campaña objetivo en la cual se van a utilizar los clusters.

Como nos interesa enfocarnos en aquellos casos que sean de bajo riesgo se incorpora la columna categórica *“credit\_risk”* y adicionalmente se incorpora la variable *“personal\_status\_sex*” para diferenciar futuras campañas por sexo.



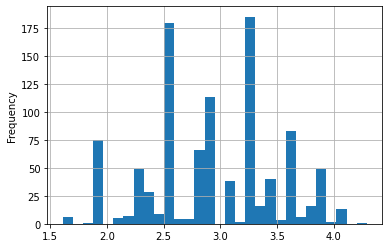
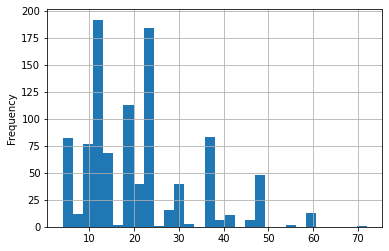
* **Categóricas a Dummies:** La variable “*credit\_risk*” no fue necesaria cambiarla ya que viene en formato 1 y 0, en cambio la variable “*personal\_status\_sex”* se le da tratamiento para agrupar en solo 2 categorías (0: Hombres y 1:Mujer) .

Se crea la variable sexo, a partir de las categorias de “*personal\_status\_sex”*

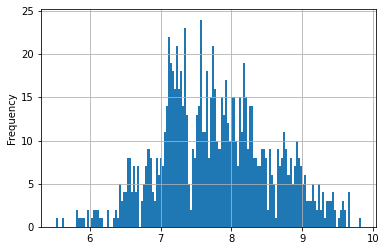
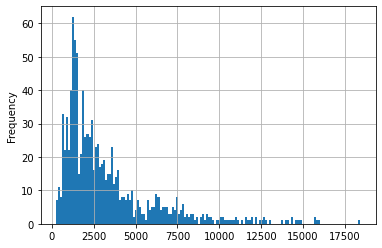


* **Transformación Cox-Box**: Las variables amount y duration resultaron ser asimétricas por lo tanto se procede a normalizar la data aplicando logaritmo natural.

Gráfica “duration” con datos originales y normalizados:



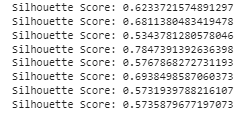
Gráfica “amount” con datos originales y normalizados:

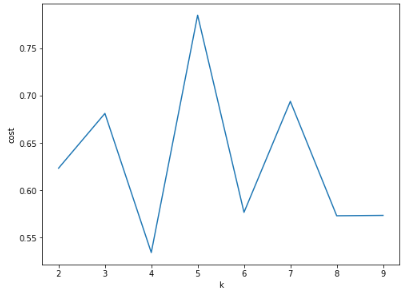


* **Normalización**: Por último las 5 variables seleccionadas son normalizadas en escala 0-1 para que todas tengan el mismo peso dentro del algoritmo esto es necesario.

B) Cómo se determinó el número óptimo de clusters a utilizar

Se aplica el método de silueta, en donde, el valor de la silueta es una medida de cuán similar es un objeto a su propio cúmulo (cohesión) en comparación con otros cúmulos (separación), es decir a mayor valor de un conjunto de clusters simulados, mejor será la segmentación.

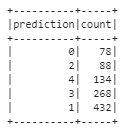




Como se observa en la gráfica el mayor valor fue de **0.78** con una cantidad de **5 clusters**, lo cual nos indica que es la mejor cantidad de grupos que se pueden obtener con estos datos.

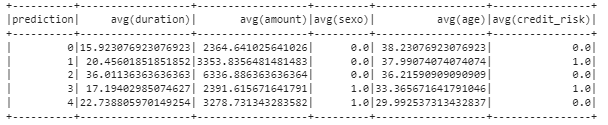
C) La estadística descriptiva de los segmentos encontrados y qué nombre “comercial” le pondría al segmento dadas dichas características

Lo primero que observamos es la cantidad de clientes que se tienen por cada segmento, es habitual que que al realizar segmentación con k-medias se esperen grupo homogéneos en cantidad, pero en este caso como estamos en presencia de variables dummies es posible que esto no se cumpla.



Al verificar los centroides o promedios de cada segmento se observa que los grupos [0,1,2] corresponden a los Hombres, y los grupos [3, 4] a mujeres.

Por otra parte sólo hay 2 grupos sin riesgo (*credit\_risk*=1) que son el grupo 1 y el grupo 3, que serán aquellos a los cuales se van a enfocar las acciones de negocios.

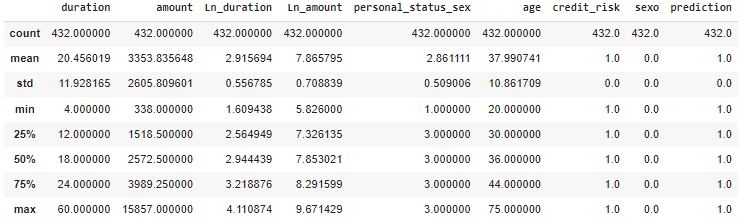


D) Acciones de negocios que podrían ser relevantes a sugerir para los segmentos encontrados

Las acciones de negocio irán enfocadas en 2 segmentos o grupos objetivos, que son los hombres sin riesgo y mujeres sin riesgo cada grupo tiene sus propias características que se verán a continuación:

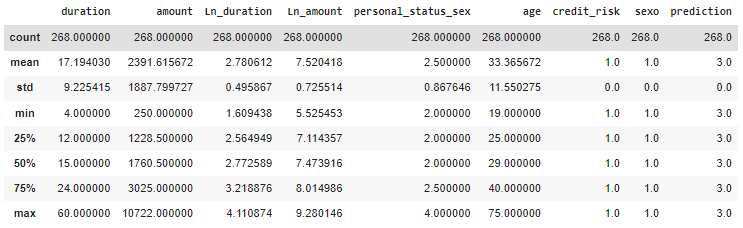
* El primer grupo de acción es el de Hombres sin riesgo, en donde solo se considerarán aquellos casos que estén en el 50% más poblado de la muestra, es decir, entre el percentil 25 y 75.

Se consideran en la edad de entre 30 y 44 años, para créditos con una duración entre 12 y 24 meses con montos entre 1518 DM y 3989 DM.



* El segundo grupo de acción es el de Mujeres sin riesgo, en donde solo se considerarán aquellos casos que estén en el 50% más poblado de la muestra, es decir, entre el percentil 25 y 75.

Se consideran en la edad de entre 25 y 40 años, para créditos con una duración entre 12 y 24 meses con montos entre 1228 DM y 3025 DM.



**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.