# Objetivo

Practicar con la carga de ficheros.

Practicar con la limpieza de los datos, substitución de campos, creación de campos nuevos.

Practicar con los joins para mejorar la calidad y ver como diseñar en la capa plata para mejorar rendimiento.

Practicar con los visores de gráficos de databricks.

# Montaje.

El 90% del código es similar a Spark, por lo que la carga, la limpieza, los joins y la construcción de ficheros y capas se realizara el GitHub CodeSpace ya que nos va a proporcionar un entorno para poder ejecutar y practicar.

Para ello durante el ejercicio iré proporcionando videos para explicar como ir montando todo.

Antes de iniciar el ejercicio hay que montar ese CodeSpace en GitHub siguiendo el video **01 Instalacion CodeSpace**

# Genérico:

Eliminar nombres en inglés

Eliminar cualquier campo nulo, con alguna estrategia de reemplazo.

# De Ficheros a Bronce.

## Carga de Ficheros.

Ver 02 CargaFicherosEjercicio en la carpeta Videos.

## Products.(json - Array)

MakeFlag, se lee como cadena y luego se transformarán a booleanos.

* Quitar:
  + ProductNumber
  + FinishedGoodFlag,
  + SafetyStockLevel
  + ReorderPoint
  + StandardCost,
  + SizeUnitMeasureCode
  + WeightUnitMeasureCode
  + DaysToManufacture
  + ProductLine
  + ProductModelID,
  + DaysToManufacture
  + Rowguid
  + ModifiedDate
  + DiscontinuedDate
* En la introducción de datos, un señor que era muy despistado, ha medido los colores en Español, afortunadamente sólo introducía valores de productos: plata (Silver), rojos (Red) y Ambar (Yellow), por lo que cambiaremos estos colores al inglés, el color null es que es transparente y le pondremos “Sin Color” cómo Color.
* En Size si es nulo pondremos (Sin Tamaño)
* Si Weight es nulo pondremos 0.
* Si Class es nulo pondremos SC (Sin Clase)
* Si Style es nulo pondremos SE (Sin Estilo)
* En subcategoria si es null pondremos -1 (Una subcategoría que añadiremos a mano)

## ProductSubCategory (JSON array)

* Quitar Rowguid
* Quitar ModifiedDate

## ProductCategory (el fichero no tiene cabecera) CSV y separado por ;

* Tomar los campos ProductCategoryID y Name y Quitar los demás.

## Customers, Clientes (CSV)

* Existen CustomerID que tienen como PersonID nulo. Eso quiere decir que no son personas, sino sistemas automáticos de compras. Pondremos en ese personID el valor -1, para indicar que no ha sido informado.
* Las columnas AccountNumber, rowguid, store y ModifiedDate deben ser eliminadas.

## Person (JSON)

* Crear un campo nombre completo. Conformado por Title, FirstName MiddleName y LastName, eliminar los campos origen.
* Eliminar las columnas NameStyle, Suffix, AdditionalContactInfo, Demographics Rowguid y EmailPromotion
* Añadir -1 para persona no informada y -2 para persona no encontrada.

## Sales Detail (CSV)

* Habrá ProductID que estarán en null, estos son líneas de pedido que no contienen ningún producto (lo pondremos como producto no informado), ProductID = -1
* El Line Total se ha metido con puntos y no se puede reconocer, transformarlo a tipo decimal, multiplicando la cantidad por el precio.
* Eliminar Rowguid, Carrier Tracking, SpecialOrderID

## SalesOrderHeader (CSV)

* Habrá CustomerID que tengan NULL eso significa que el cliente ha sido interno. Por ejemplo no consideramos clientes a los propios trabajadores. Lo pondremos como -1.
* Habrá CustomerID que no tengan un código válido, eso es porque se han borrado los código de los clientes, se podrá como -2
* Eliminar AccountNumber, SalesPersonID, RevisionNumber, OnlineOrder, SalesOrderNumber, PurchaseOrderNumber, BillToAddressID, ShipToAddressID, CreditCardApprovalCode, CurrencyRateID, Subtotal, TaxAmt, Freight, TotalDue, Comment, Rowguid, CreditCardID,ModifiedDate
* Las fechas las quiero convertidas en un número entero, me explico “24/03/2033” debería de ser 20330324, hacerlo para la fecha del pedido, la fecha de vencimiento y la fecha de envío.

## Vendedores: SalesPerson.csv

Tenemos que leer la tabla Sales Person y relacionarla con Person para conocer su nombre, eliminar Rowguid, SalesYTD y SalesLastYear. La fecha ponerla como Entero.

# De bronce a Plata

Vamos a desnormalizar.

## DimProducto.

* Product,
* ProductSubCategory,
* ProductCategory
* Los campos que tiene que tener DimProducto quedará como sigue:
* ProductID
* NombreProducto
* PrecioCatalogo
* Tamaño
* Peso
* Clase
* Estilo
* Color
* EnProd
* NombreSubCategoria
* NombreCategoria

## DimCliente

* Cursomers
* Person
* Los campos que tiene que tener DimProducto quedará como sigue:
* CustomerID
* TerritoryID
* PersonType
* Nombre

## DimFechas:

* Creada a mano.

dateInt

CalendarDate| Ano|NombreMes|Mes|NombreDia|DiaDeLaSemana|DiaDelMes|EsUltimoDiaDeMes|Dia|Semana|Trimestre|

Tabla

Descripción generada automáticamente

## FactVentas.

* SalesOrderDetail
* SalesOrderHeader
* Los campos que tiene que tener FactVentas, quedará como sigue:
* SalesOrderID
* SalesOrderDetailID
* Cantidad
* ProductID
* Precio
* TotalLinea
* UnitPriceDiscount
* Status
* CustomerID
* TerritoryID
* ShipMethodID
* FechaPedido
* FechaVenc
* FechaEnv|

## DimVendedores

* SalesPerson
* Person
* Los campos que tiene que tener DimVendedor quedará como sigue:
* SalesPersonID
* TerritoryID
* PersonType
* Nombre

# De plata a Oro

NOTA: Intentar hacerlo re

El departamento de Ventas y el gerente nos piden unos resúmenes de este tipo.

Resumen de venta por cliente, la granularidad que nos solicitan es por mes y año.

* VClienteAnoMes

Resumen de venta por territoryID, Mes y Año

* VTerroryID

El departamento de Producto, nos piden:

Resumen de venta por producto, Mes y Año.

* VProductoAnoMes

Resumen de venta por categoría, subcategoria, Mes y Año

* VProductoCategoriaSubcategoriaMesAno

El departamento de Marketing quiere saber para cada cliente, cuanto ha comprado por Categoria y Año.

* VMarketingCategoriaAno

El departamento de Ventas, quiere saber para cada vendedor cuanto vendido de cada categoría de producto.

* VVendedoresCategoria
* También quiere ver cada vendedor a que cliente ha vendido y acumular el total por años.

# Informes, gráficos y Dashboards.

## Gráfico de Barras de Pedidos, métodos de envio.

**Valores de configuración**: para esta visualización del gráfico de barras, se establecieron los siguientes valores:

* Columna X:
  + Columna del conjunto de datos: OrderDate
  + Nivel de fecha: Months
* Columnas Y:
  + Columna del conjunto de datos: totalOrderAmount
  + Tipo de agregación: Sum
* Agrupar por (columna del conjunto de datos): shipMethod
* Apilamiento: Stack
* Nombre del eje X (invalidar el valor predeterminado): Order Month
* Nombre del eje Y (invalidar el valor predeterminado): TotalOrderAmount

## Gráfico de líneas

**Valores de configuración:** para esta visualización del gráfico de líneas, se establecieron los siguientes valores:

* Columna X:
  + Columna del conjunto de datos: OrderDate
  + Nivel de fecha: Years
* Columnas Y:
  + Columna del conjunto de datos: totalOrder
  + Tipo de agregación: Average
* Agrupar por (columna del conjunto de datos): ShipMethod
* Nombre del eje X (invalidar el valor predeterminado): Order year
* Nombre del eje Y (invalidar el valor predeterminado): Average price

## Gráfico de áreas

**Valores de configuración:** para esta visualización del gráfico de áreas, se establecieron los siguientes valores:

* Columna X:
  + Columna del conjunto de datos: OrderDate
  + Nivel de fecha: Years
* Columnas Y:
  + Columna del conjunto de datos: TotalOrder
  + Tipo de agregación: Sum
* Agrupar por (columna del conjunto de datos): o\_shipMethod
* Apilamiento: Stack
* Nombre del eje X (invalidar el valor predeterminado): Order year
* Nombre del eje Y (invalidar el valor predeterminado): Total Order

## Gráficos circulares

**Valores de configuración:** para esta visualización del gráfico circular, se establecieron los siguientes valores:

* Columna X (columna del conjunto de datos): o\_shipMethod
* Columnas Y:
  + Columna del conjunto de datos: o\_totalOrder
  + Tipo de agregación: Sum
* Etiqueta (invalidar el valor predeterminado): Total Order

## Histogramas

**Valores de configuración:** para esta visualización del histograma, se establecieron los siguientes valores:

* Columna X (columna del conjunto de datos): o\_totalAmount
* Número de cubos: 20
* Nombre del eje X (invalidar el valor predeterminado): Total Amount

## Mapas térmicos

**Valores de configuración**: para esta visualización de los mapas térmicos, se establecieron los siguientes valores:

* Columna X (columna del conjunto de datos): ship\_Method
* Columnas Y (columna del conjunto de datos): o\_orderstatus
* Columna de color:
  + Columna del conjunto de datos: o\_totalAmount
  + Tipo de agregación: Average
* Nombre del eje X (invalidar el valor predeterminado): ship\_Method
* Nombre del eje Y (invalidar el valor predeterminado): TerritoryID
* Combinación de colores (invalidar el valor predeterminado): YIGnBu

## Gráfico de dispersión

**Valores de configuración:** para esta visualización del gráfico de dispersión, se establecieron los siguientes valores:

* Columna X (columna del conjunto de datos): cantidad
* Columna Y (columna del conjunto de datos): price
* Agrupar por (columna del conjunto de datos): ProductID
* Nombre del eje X (invalidar el valor predeterminado): cantidad
* Nombre del eje Y (invalidar el valor predeterminado): Price

## Gráfico combinado

**Valores de configuración:** para esta visualización de gráfico combinado, se establecieron los siguientes valores:

* Columna X (columna del conjunto de datos): FechaEnvio
* Columnas Y:
  + Primera columna del conjunto de datos: Precio
  + Tipo de agregación: promedio
  + Segunda columna de conjunto de datos: Cantidad
  + Tipo de agregación: promedio
* Nombre del eje X (invalidar el valor predeterminado): FechaEnvio
* Nombre del eje Y izquierdo (invalidar el valor predeterminado): Cantidad
* Nombre del eje Y derecho (invalidar el valor predeterminado): Average Precio
* Series:
  + Order1 (columna de conjunto de datos): AVG(Precio)
  + Eje Y: derecho
  + Tipo: Línea
  + Order2 (columna del conjunto de datos): AVG(Cantidad)
  + Eje Y: izquierdo
  + Tipo: barra

# Acceso a información fuera de Databricks (en cuenta de almacenamiento externa).