**Prueba técnica Data Engineer**

**Mercado Libre**

**Nombre**: Juan David Espitia Aguillon

**Objetivo:**

Esta prueba consta de tres puntos de desarrollo, cada uno con fines y objetivos diferentes

con los cuales buscamos poner a prueba sus habilidades técnicas en diseño y

programación.

**Stack Tecnológico empleado para la prueba:**

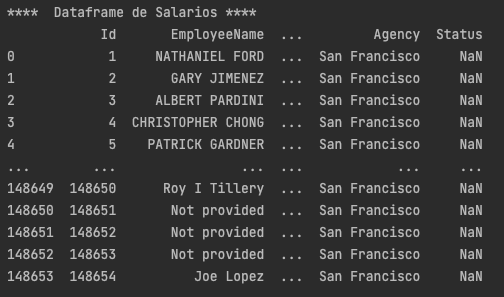
1. Lenguaje de programación: Pyhon 3.9
2. Control de Versiones: Git + GitHub + GitFlow
3. Framework: PyCharm
4. Diseños: Draw.io
5. Nube: Google Cloud Platform

**Punto 1:**

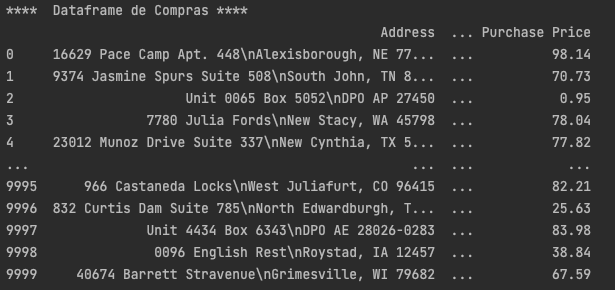
1. De acuerdo al set de datos SALARIOS ¿Cuantos cargos estaban ocupados solamente por una persona en 2011?

Inicialmente cargamos los set de datos a dataframes con la librería de Pandas

Dataframe de Salarios



Dataframe de Compras

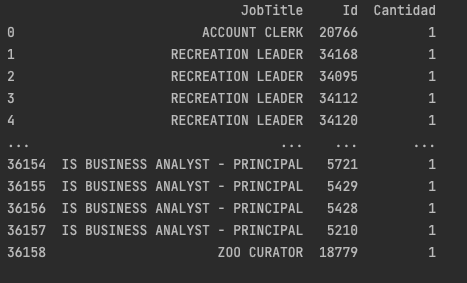


Imprimos en pantalla la cantidad de registros por cada uno



Filtramos el set de datos de salarios para que solo esten los datos pertenecientes al año 2011.

Luego agrupamos el set de datos con los campo JobTitle e ID y obtenemos la cantidad de cargos ocupados por una única persona y posteriormente nos quedamos con aquellos cuyo conteo haya sido de uno solo.

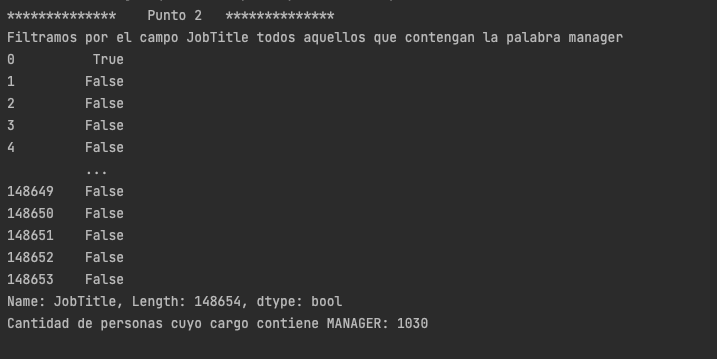


La cantidad final es de:



1. De acuerdo al set de datos SALARIOS ¿Cuánta gente tiene la palabra 'MANAGER' en su cargo?

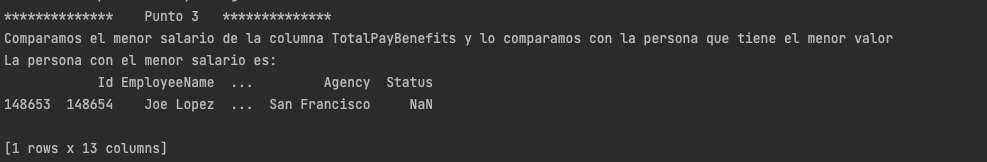
Inicialmente filtramos por el campo JobTitle todos los que contengan la pabra manager y luego los sumamos para obtener la cantidad.



La cantidad de personas con la palabra MANAGER son: 1030

1. De acuerdo al set de datos SALARIOS ¿Cuál es el nombre de la persona que menos gana (incluyendo beneficios - TotalPayBenefits)?

Se debe comparar el menor salario de la columan TotalPayBenefits con la persona que obtiene ese mismo valor

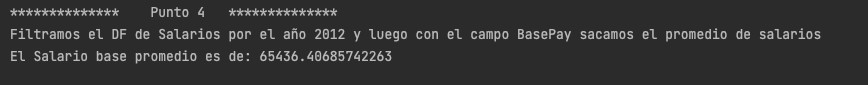


La persona con el menor salario es:

Joe Lopez ID: 148654

1. De acuerdo al set de datos SALARIOS¿Cuál es el salario base (BasePay) promedio de todos los empleados para el año (2012)?

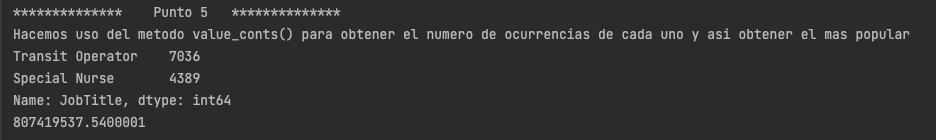
Filtramos el DF de slarios por el año 2012 y luego con el campo de BasePay (Salario Base) obtenemos el promedio.



El Promedio de salario es: 65436,407

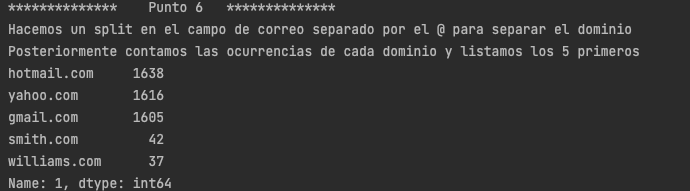
1. De acuerdo al set de datos SALARIOS ¿Cuál fue la suma total pagada con beneficios por los dos trabajos más populares?.

Hacemos uso del metodo value\_conts() para obtener el número de ocurrencias de cada uno y asi obtener el mas populares y Posteriormente usamos la sentecia para sumar el total de benefios pagados de ambos trabajos.



El total pagado por los 2 trabajos mas populares es: 807419537.540001

1. De acuerdo al set de datos COMPRAS ¿Cuáles son los 5 proveedores de correo electrónico más comunes, con cuantos usuarios está asociado cada uno? (hotmail.com,gmai.com, etc)

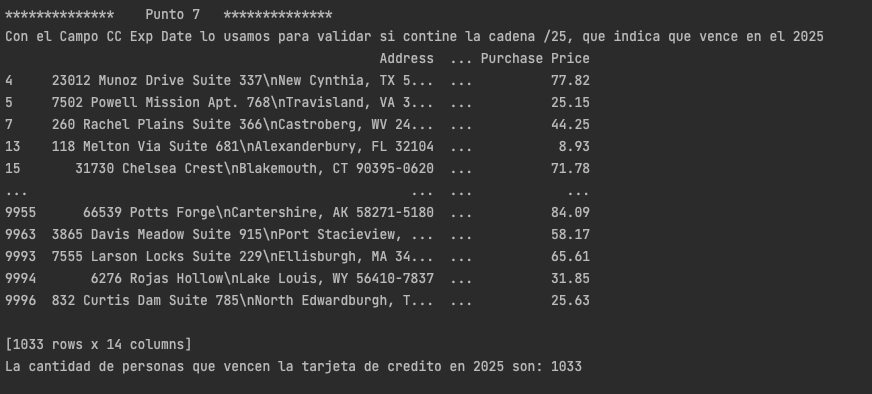


Los 5 correos mas comunes son:

|  |  |
| --- | --- |
| **Proveedor** | **Ocurrencias** |
| Hotmail.com | 1638 |
| Yahoo.com | 1616 |
| Gmail.com | 1605 |
| Smith.com | 42 |
| Williams.com | 37 |

1. De acuerdo al set de datos COMPRAS ¿Cuántas personas tienen una tarjeta de crédito que expira en 2025?

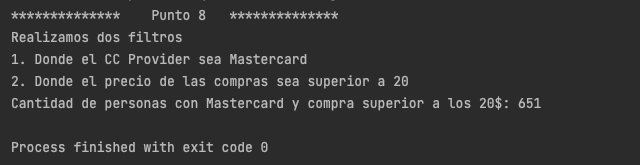
Con el Campo CC Exp Date lo usamos para validar si contine la cadena /25, que indica que vence en el 2025 y luego usamos la función len para validar la cantidad de registros reportados



La cantidad de personas que vence su tarjeta de crédito en 2025 son: 1033

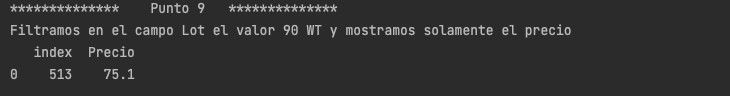
1. De acuerdo al set de datos COMPRAS ¿Cuántas personas tienen tarjetas Mastercard e hicieron una compra por más de $20?

Debemos realizar dos filtros, el primero sobre el campo CC Provider cuyo valor sea igual a Mastercard y el segundo sobre el campo Purchase Price cuyo valor sea superior a los 20$



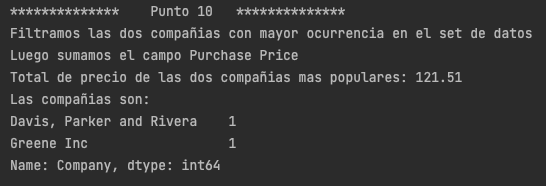
La cantidad de personas con Mastercard y compra superior a los 20$ son: 651

1. De acuerdo al set de datos COMPRAS ¿Alguien hizo una compra desde Lot: "90 WT", ¿cuál fue el precio de compra de esta transacción?



El precio final es de 75.1

1. De acuerdo al set de datos COMPRAS ¿ Cuánto suma el total de precio de compras para las dos compañías menos populares?, ¿Cuáles son esas dos compañías?



El total de precio de las dos compañías mas populares es de: 121.51 y las dos compañías son:

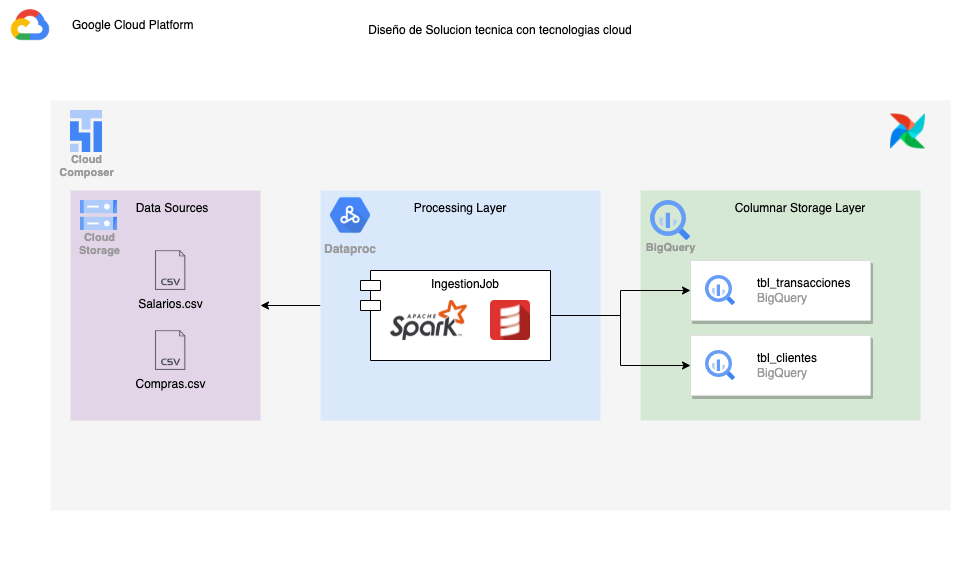
* Davis, Parker and Rivera
* Greene Inc

**Punto 2:**

**Diseño Cloud – GCP**

Diseño de la solución con tecnologías Cloud. Para este ejercicio se uso la nube de Google con sus servicios para procesar datos.

La arquitectura general de solución es la siguiente:



**Nota:**

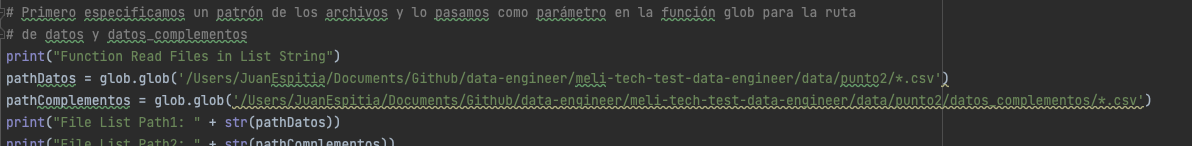
El diseño está pensado en un pipeline de datos con un alto volumen, por tal razon se necesita de servicios idóneos para la transformación de los datos.

El Componente de IngestionJob, sería un desarrollo construido en el framework de procesamiento de datos Apache Spark, bajo el lenguaje de programación Scala, donde se implementaría la lógica de lectura de los archivos, las transformación de concatenación, hash, homologación y persistencia de los datasets finales en BQ  
  
Si se piensa un diseño con set de datos de menor volumen, podemos reemplazar en la capa de procesamiento por el artefacto de Scala por una Cloud Function, desarrollada con Python y podemos utilizar el código desarrollado para la prueba en el punto 2.

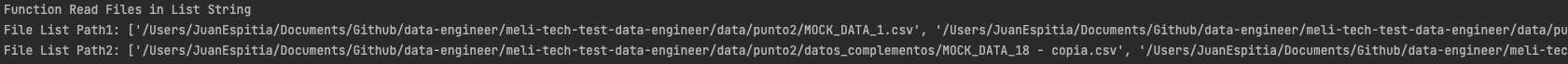
El desarrollo se realizó de la necesidad se implemento en lenguaje Python por efectos de agilidad con la entrega de la prueba técnica

**Implementación**

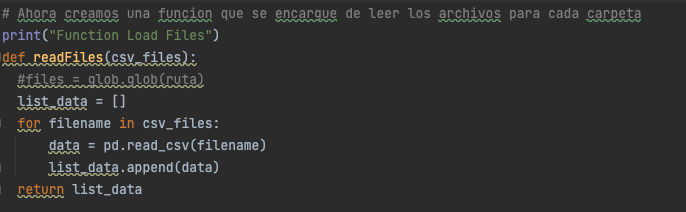
El primer paso es establecer las rutas donde se encuentran los datos a ser cargados.

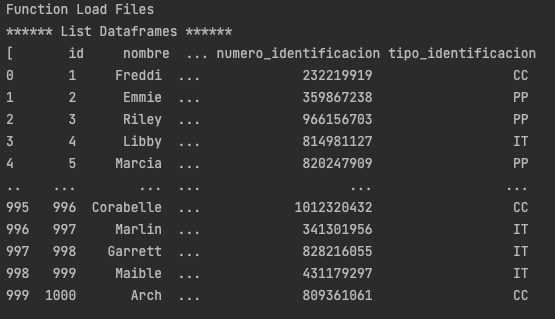


Resultado:

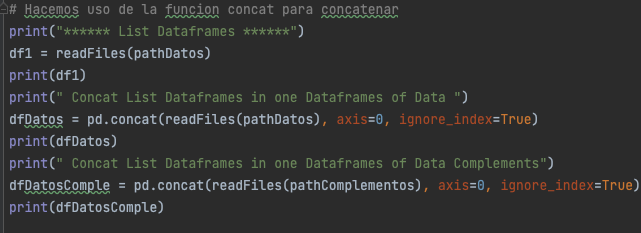


Seguido, creamos una función que itere cada una de las rutas con sus respectivos archivos .csv y vamos cargando uno a uno de los archivos .csv en una lista de dataframes y retornamos la lista de lista de dataframes.

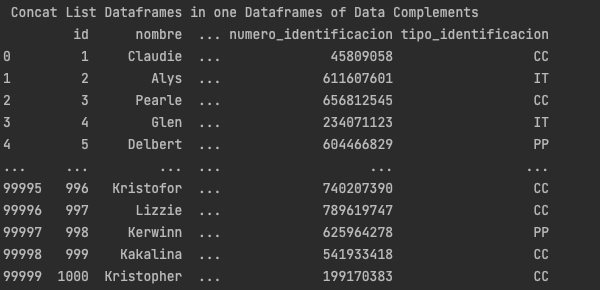




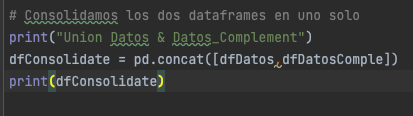
Luego hacemos uso de la función de Python concat() de dataframes para concatenar los elementos que contiene la lista que nos retorna la función anterior y lo convertimos en un dataframe para cada ruta de datos.



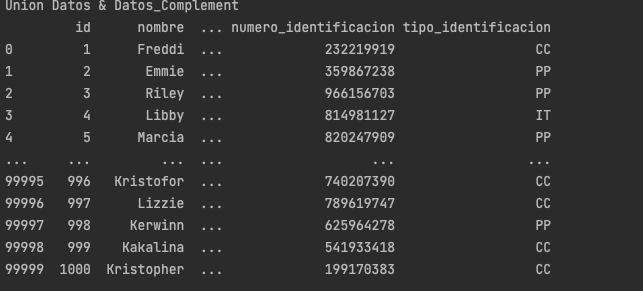
Output



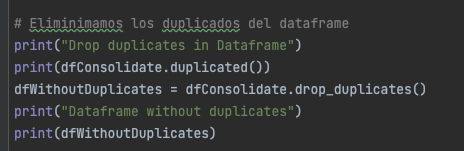
Luego consolidamos los dataframes de datos y datos\_complementarios en un solo dataframe, nuevamente con la función concat()



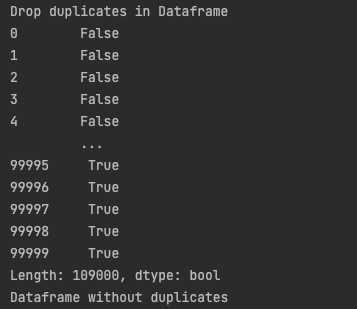
Output



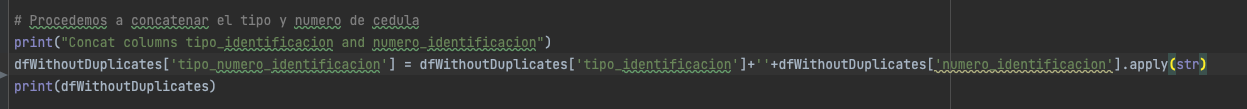
Posteriormente procedemos a eliminar los registros duplicados del dataframe consolidado mediante la función de drop\_duplicates()



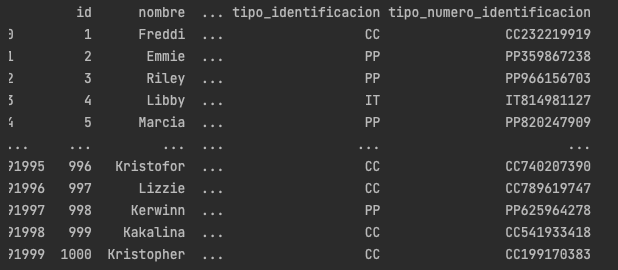
Output



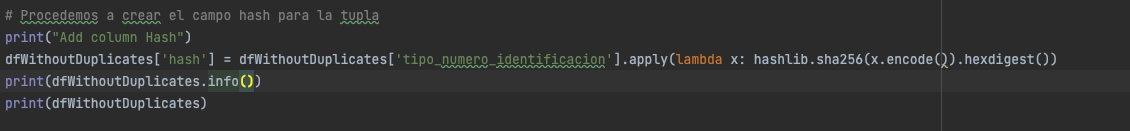
Luego procedemos a concatenar los campos tipo y numero de documento mediante las siguientes sentencias de código



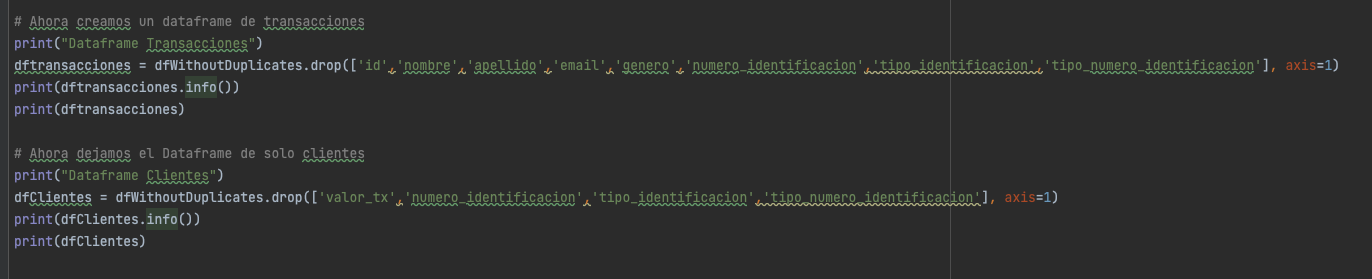
Y obtenemos la siguiente salida



Posterior a la concatenación de columnas realizamos el proceso de encriptar la columna genreda de la concatenación de tipo y numero de documento, esto mediante las siguientes sentecias.



Ahora separamos los dos set de datos requeridos, el de Clientes y el de Transacciones con las especificaciónes de campos indicados en el documento de la prueba técnica.

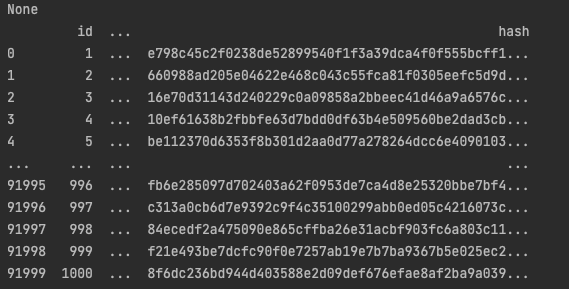


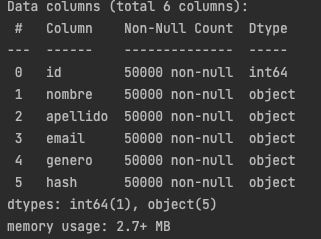
Dataframe transacciones



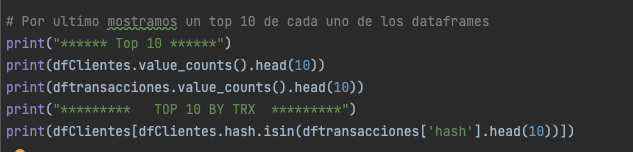
Dataframe de Clientes

Nota: No se visualizan bien todas las columnas





Por ultimo mostramos el Top 10 de los clientes con mas transacciones realizadas



Salida por consola

