**Nombre: Julián Andrés González Vélez**

**CC: 1116.237.362**

**Email:** [**j.andres2087@gmail.com**](mailto:j.andres2087@gmail.com)

**Respuestas a la prueba técnica para analista de ciencia de datos**

1 Datos:

**Tipos/Variables de datos:**

* **Cuantitativos:** Aquellos que se pueden representar en tipos de datos numéricos (enteros, decimales, complejos), por lo cual se pueden realizar operaciones matemáticas sobre estas. Como se mide?
* **Cualitativos:** Son aquellas que representanun atributo o cualidad propia del sujeto a evaluar, por lo general son cadenas de texto o caracteres que representan categorías o estado. .Cómo se mide?

**-Estructuras de datos en pyhton**

**Xxxxxx**

**-Tipos de bases de datos**

**Xxxxx**

**2) 4Vs Big Data**

**Volume:** Hace referencia a las diferentes fuentes desde donde se generan datos, haciendo que la cantidad hoy en día sea super grande. Lo anterior nos lleva a que debemos tener en cuenta la cantidad de información a evaluar ya que al venir de fuentes como redes sociales, dispositivos de los usuarios, automáticos entre otros, nos pone retos interesantes en temas de procesamiento de estos datos como por ejemplo herramientas de clusterización del cómputo, mejoramiento del performance entre otros.

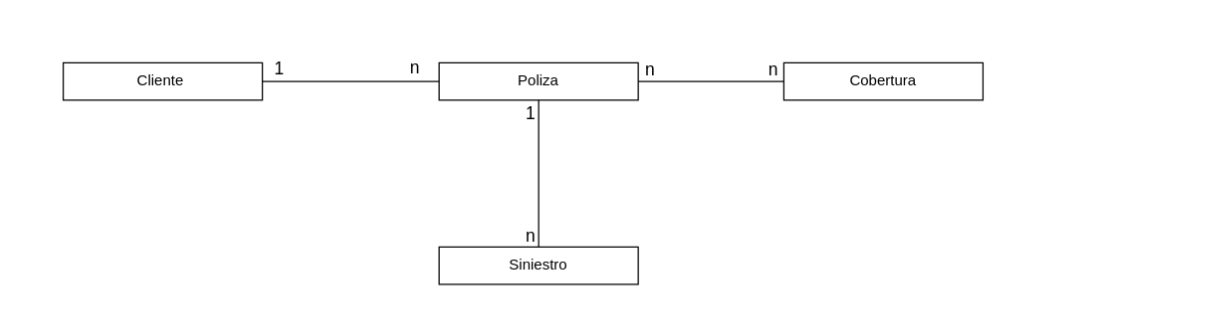
**Variety:** Hace referencia a que la información que obtenemos de las diferentes fuentes puede ser diferentes en su estructura, por ejemplo, hoy en día nos encontramos información en videos, imágenes, correos, tweets y note que son fuentes muy distintas a las de antes que por lo regular son bases de datos. Big Data establece lo anterior como datos estructurados (bases de datos, json u otra estructura más intuitiva de leer) y no estructurados, este último nos obliga a pensar un poco más del como vamos a procesar los datos y poder inferir u obtener información relevante.

**Velocity:** Hace referencia a la frecuencia en que los datos deben ser evaluados, ya que cambian muy rápido y no queremos que estemos pintando información desactualizada, aquí es donde el procesamiento streaming de los datos es relevante para estar actualizados en tiempo real.

**Veracity:** Hace referencia a que podamos establecer que información es valiosa y que información genera ruido y son datos basura y así poder generar información confiable.

Todas las anteriores son útiles para obtener inferencias/conclusiones válidas acerca de la muestra(datos) que se están analizando.

**3 Diagrama entidad relación:**



**4 Nube**

**Características principales:**

* Ofrecen múltiples servicios en cuanto a análisis de datos**,** infraestructura, desarrollo de software**.**
* Modalidad de pay as you go, esto es que permite configurar el pago por consumo de servicios, por ejemplo, que se pagará por el tiempo en el que una tarea se ejecuta en la nube y al terminar de ejecutarse se apagará el recurso o servicio y no se cobrará por el tiempo de inactividad.
* Apuntan a que se merme la administración de los servicios, ya que estos no hacen parte de la compañía y en su lugar se transfiere al proveedor de nube.

**Elasticidad y escalabilidad en la nube:**

La nube ofrece escalamiento tanto horizontal como vertical, al permitir ampliar la capacidad de computo adicionando nodos y/o capacidades de máquina (cambiando de especificaciones de mi máquina en la nube)

**Elasticidad**

Gestión de infraestructura en cuestión de segundos, ya que podemos escalar fácilmente para eventos picos que se tengan, por ejemplo en temporadas altas en que el sistema va a ser más accedido entonces se puede levantar más nodos rápidamente y el performance de nuestro sitio aumentaría y daría cabida a mayor peticiones concurrentes.

**Capas de la computación en la nube:**

* **SAAS:**
* **PAAS:**
* **IAAS:**

**5 Principales procesos de limpieza de datos**

**6 Formas de normalización de una Bd relacional**

**1FN (Primera forma normal):** En ninguna columna podemos encontrar múltiples valores (separados por coma por ejemplo) y no hay grupos repetidos.

**¿Cómo logarlo?:** Cree una nueva tabla que represente la columna que tiene valores múltiples y cada valor en la nueva tabla debe tener su id único y debe además cumplir obviamente con la 1FN

**2FN (Segunda forma normal**): Debe cumplir la 1FN

**¿Cómo logarlo?:**

**3FN (Tercera forma normal):** Debe cumplir la 2FN

**¿Cómo logarlo?:**

**7 Scripts de Python** (ver documento suta\_test.py)

**8 Función Python** (ver documento suta\_test.py)