**Analítica de Grandes Datos**

**Departamento de Ciencias de la Computación y la Decisión**

**Facultad de Minas**

**Universidad Nacional de Colombia**

Trabajo Nociones de Arquitectura de la Información

Versión: aaaa.mm.dd HH:mm

*Observación: Cada vez que agregue nuevos elementos al documento, o que modifique algún componente del informe, revise la coherencia y consistencia con los otros elementos que hacen parte del mismo.*

**Responsables**

|  |
| --- |
| Nombre Completo – Documento de Identificación |
| 1. |
| 2. |
| 3. |
| 4. |
| 5. |
| **REPO EN GITHUB:** |

*Realiza este trabajo considerando los datos que generan los sistemas transaccionales e información no estructurada de tu dominio (si trabajas por ejemplo para TCC tu dominio es la mensajería; también puedes explorar en la página* [*https://www.kaggle.com/datasets*](https://www.kaggle.com/datasets) *o* [*https://arxiv.org/*](https://arxiv.org/)*). Considera tener acceso a esta información, de al menos 10 MB (puede ser uno o varios archivos de texto), y* ***tener al menos cuatro clases conceptuales. Este documento también debe almacenarse en el REPO. Plazo Máximo de Entrega 23 de Mayo, NO SE recibirá por correo electrónico, envío por*** [***https://forms.gle/h7ty3yZykaUq5m7y6***](https://forms.gle/h7ty3yZykaUq5m7y6)

1. **Comprensión del negocio**
   1. **Descripción del contexto del negocio.**

Describa en máximo 250 palabras el contexto en el cual se generan los datos y cuál es el proceso que los genera.

El contexto de los datos está soportado en la Resolución 3374 de 2000, por la cual se reglamentan los datos básicos que deben ser reportados por los prestadores de servicios de salud y las entidades Administradoras de Planes de Beneficios, la información del Registro Individual de Prestaciones de Servicios de Salud - RIPS, constituyen una fuente de datos prioritaria para los procesos de dirección, regulación y control del Sistema General de Seguridad Social en Salud – SGSSS.

El RIPS provee los datos mínimos y básicos, y facilita las relaciones entre las entidades administradoras de planes de beneficios (pagadores), las instituciones prestadoras de servicios de salud y los profesionales independientes (prestadores) apoyando la generación de planes, programas, estrategias y operaciones, para la toma decisiones gerenciales, estratégicas, administrativas y asistenciales.

Se define analizar los registros individuales para el servicio de hospitalización en el municipio de Medellín, entre los años 2009 a 2017.

* 1. **Identificación del problema:**

Delimite en máximo 150 palabras la problemática, así como identificar los requisitos, supuestos, restricciones y beneficios de la solución de este.

Conocer la frecuencia de utilización de los servicios de hospitalización en el municipio de Medellín, la morbilidad y su caracterización según variables sociodemográficas, con el fin de determinar la demanda de los servicios de hospitalización y permita destinar recursos en el fortalecimiento de la red hospitalaria (generar políticas en salud)

Requisitos:

Supuestos:

Restricciones: Sin embargo, y pese a que los distintos generadores y usuarios de los datos reconocen su valor e importancia, en la actualidad se han identificado falencias tanto en la calidad del registro primario, como en la utilización para la gestión de las IPS y las EAPB, así como para la formulación de la política pública en salud.

Beneficios de la solución: Definir políticas públicas a morbilidad y su caracterización según variables sociodemográficas, así mismo contribuir a generar guías de práctica clínica para optimizar el cuidado del paciente, disminuir la variabilidad innecesaria en la práctica clínica.

**Determinación de objetivos:**

Describa en máximo 150 palabras las metas a lograr al proponer una solución basada en un modelo de datos o de analítica (cómo y qué tipo de ventaja competitiva se ganará).

* 1. **Evaluación de la situación actual:**

Describa en máximo 150 palabras el estado actual antes de implementar la solución de analítica, a fin de tener un punto de comparación que permita medir el grado de éxito de la solución.

1. **Comprensión de los datos**
   1. **recolección de datos**

Describa en máximo 150 palabras los datos a utilizar identificando las fuentes, las técnicas empleadas en su recolección, los problemas encontrados en su obtención y la forma como se resolvieron los mismos. Además, adjunte los datos (archivos de texto, etc.) agréguelos en el github **(REPO EN GITHUB)** en un solo archivo, por favor comprímalo(s). Llame el archivo T1.2.1.Datos.zip

* 1. **Descripción de datos (diccionario):**

Diligencia la siguiente tabla, puede agregar otra columna si lo considera necesario.

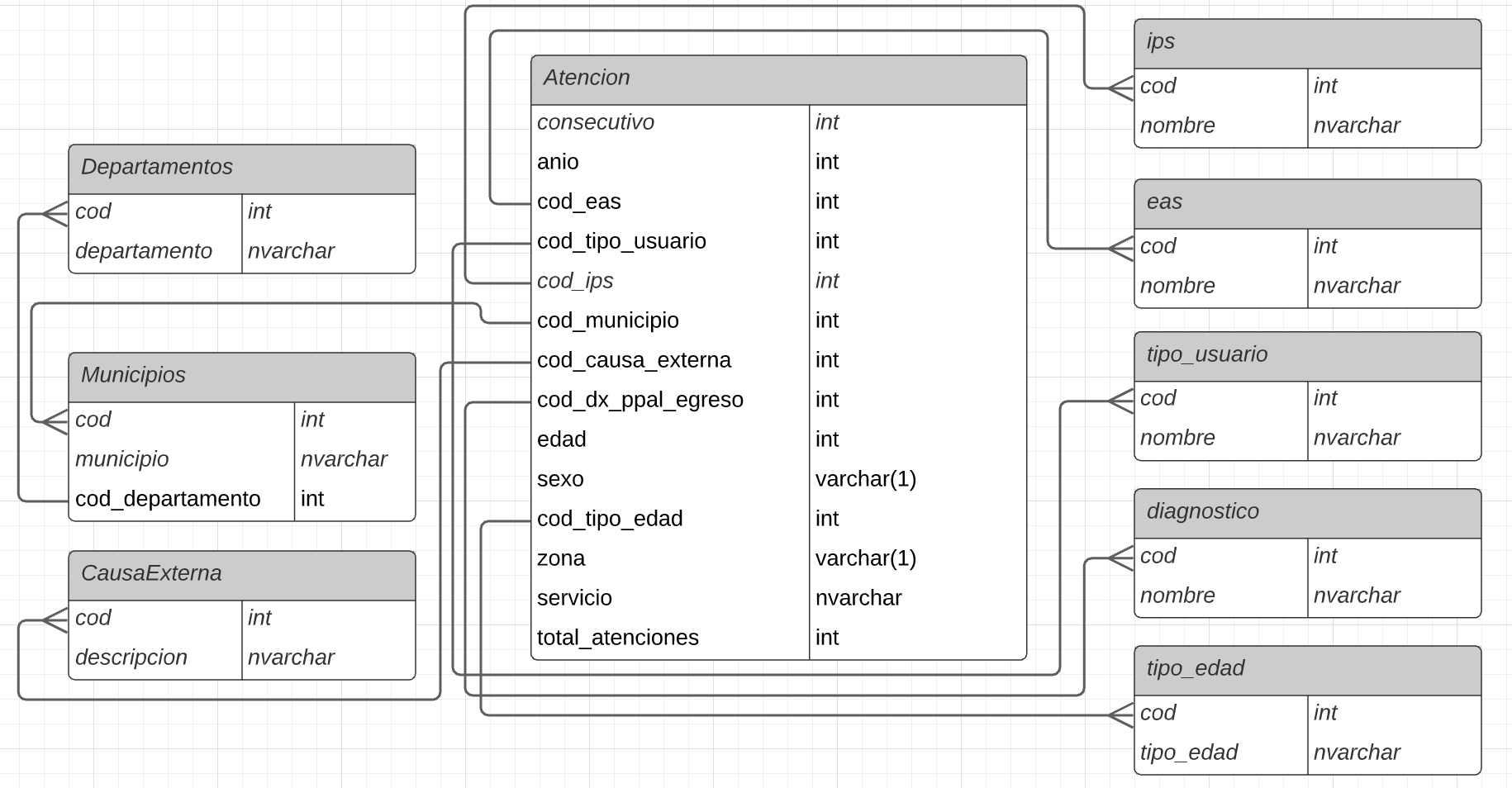
|  |  |  |
| --- | --- | --- |
| Nombre del atributo / variable | Formato o Tipo de Dato | Descripción |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

* 1. **Modelo del dominio**

***Observación****: Incluya el gráfico del modelo del dominio que representa la estructura de datos de su problema.*

1. **Modelo Entidad-Relación**
   1. **Toma de pantalla del modelo E-R**

***Observación****: lo que se pide, puede usar https://draw.io o Microsoft Visio® y modele usando la notación de Barker.*

**

* 1. **Sentencia o consulta de creación del tabla(s)**

***Observación****: Escriba el código en el Sistema de Gestión de Bases de Datos Relacionales de su elección (se recomienda SQLite por simplicidad, mediante https://sqlitebrowser.org/) para crear las tablas que corresponda con su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.2.Creacion\_Tablas.sql*

***Se recomienda repasar SQL en https://www.w3schools.com/sql/default.asp***

* 1. **Sentencias para Insertar datos**

***Observación****: Escriba el código para insertar los datos en cada una de las tablas creadas. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.3.Insertar\_Datos.sql*

* 1. **Sentencia de consulta**

***Observación****: realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de pantalla donde evidencie la sentencia SQL y su correspondiente ejecución. Además, *Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.3.4.Consultar\_Datos.sql*

1. **MongoDB**
   1. **Sentencia o consulta de creación del documento(s)**

***Observación****: Escriba el código en MongoDB para crear al menos 20 documentos que correspondan a su conjunto de datos específico. Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.4.1.Creacion\_Documentos.sql*

* 1. **Sentencia de consulta**

***Observación****: Realice la exploración básica de los datos, conteos totales y por categorías, máximos, promedio y mínimos.* Es decir, aplique estadística descriptiva con el fin de conocer las propiedades de los datos y entenderlos lo mejor posible. Use solamente sentencias SQL. Anexe las tomas de pantalla donde evidencie la sentencia SQL y su correspondiente ejecución. Además, *Almacene en el repositorio* **(REPO EN GITHUB)** *el script con el nombre de T1.4.2.Consultar\_Datos.sql*

1. **Análisis de lectura**

***Observación****: Considerando el artículo: “*The Definitive Guide to Graph Databases for the RDBMS Developer*” de Neo4J. Compartido en las carpeta de lecturas recomendadas. Analice y responda cada pregunta en máximo 150 palabras:*

1. ¿Cuáles son las limitaciones, que se pueden inferir de la lectura, para migrar los conjuntos de datos relacionales a NoSQL?
2. ¿Cuáles limitaciones adicionales que se deben considerar, a parte de las mencionadas en el artículo?
3. ¿Cuáles son las razones (criterios) que se deben considerar para migrar un conjunto de datos relacionados a NoSQL?