**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: 2021.05.06 20:00

*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. Juliana Arias Ciro - 1038409725 |
| 2. Juan Pablo López Buitrago - 1037975877 |
| 3. Federico Cardona Salazar - 1053870065 |
| 4. Jhon Edwin Mejia Leon - 1111204157 |
| 5. Jorge Iván Gómez Restrepo - 9770450 |
| **REPO EN GITHUB: https://github.com/George010823/TrabajoAGD_Nutricion.git** |

*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 o 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***

# 1 Comprensión del negocio

## 1.1 Descripción del contexto del negocio.

*Dada la imposibilidad de utilizar datos particulares, el presente trabajo hace uso de una de las bases de datos contenidas en el portal de datos de Medellín (MEData). Específicamente se utiliza la Base de datos de Seguimiento Nutricional Infantil (SENIN), consolidada a partir de los reportes enviados por las IPS notificadoras que tienen el programa de crecimiento y desarrollo. La base de datos cuenta con 386.608 registros de niños menores de 6 años y está disponible en la siguiente ruta: http://medata.gov.co/dataset/estado-nutricional-de-menores-de-6-a%C3%B1os-programa-de-crecimiento-y-desarrollo*

## 1.2 Identificación del problema:

*Una buena nutrición es clave para el desarrollo de cualquier nación. En los primeros años de vida de una persona, una buena nutrición y una alimentación balanceada es fundamental tanto para su desarrollo físico, psicológico y emocional, evitando así problemas de crecimiento y disminuyendo la aparición de enfermedades y comorbilidades asociadas a una dieta con déficit calórico y nutricional. En consecuencia, no resulta extraño que en las últimas tres décadas las políticas de desarrollo a nivel mundial estén asociadas a la erradicación del hambre y la malnutrición en países en vías de desarrollo (FAO, 1993; UN, 2002), siendo en la actualidad la Agenda 2030 y los Objetivos de Desarrollo Sostenible - ODS[1] el referente más cercano en temas de políticas de desarrollo.*

*En Medellín se han dado avances significativos en materia de nutrición y erradicación del hambre en primera infancia gracias a las políticas públicas como el programa Buen Comienzo; sin embargo, aún queda camino por recorrer.*

## 1.3 Determinación de objetivos:

*La mala nutrición en la infancia representa un impedimento para el correcto desarrollo físico, emocional y cognitivo; limita la obtención de mejores condiciones de vida en el futuro y representa un obstáculo de largo plazo para mejorar la competitividad a nivel de ciudad en términos de formación de capital humano altamente cualificado. Por este motivo, el objetivo del presente trabajo es identificar las condiciones de nutrición en los niños entre 0 y 6 años de la ciudad de Medellín, y para ello se evaluará la categoría de relación peso_edad_ds, talla_edad_ds y peso_talla_ds en función de los atributos más críticos para la comuna*

[1] En total son 17 los ODS, en donde los dos primeros están asociados a la erradicación de la pobreza y del hambre a nivel mundial. Fuente: https://www.globalgoals.org/ (Consultado en: 15 de abril de 2021)

*Palmitas, dada su representatividad dentro de la encuesta y su condición de vulnerabilidad, para identificar en qué porcentaje de cada categoría se encuentra la comuna. En esencia, se busca realizar un ejercicio piloto que busque una comprobación de la eficiencia del modelo en dicho corregimiento, cuyo éxito podría contribuir a la elaboración de un modelo a nivel de ciudad.*

## 1.4 Evaluación de la situación actual:

*En la ciudad de Medellín ha habido avances significativos en las últimas dos décadas en la implementación de políticas públicas orientadas a la mitigación del hambre en la primera infancia, a través de articulaciones entre las Secretarías de Salud y Educación de la Alcaldía de Medellín y entre entidades públicas y privadas. El programa Buen Comienzo, iniciado en el año 2003, ha liderado todos los esfuerzos orientados en este sentido, y fue la fuente de inspiración para la implementación de la política a nivel nacional denominada "De cero a Siempre".*

*Contar con una base datos de nutrición infantil de fácil acceso, confiable y actualizada es fundamental para el correcto diseño de política pública e intervención temprana de niños en condiciones de riesgo nutricional o vulnerabilidad. Actualmente se cuenta con una base de datos desactualizada, pues los datos sólo aparecen hasta el año 2018, y de difícil trazabilidad.*

# 2 Comprensión de los datos

## 2.1 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**

*La base de datos utilizada para este análisis fue construida y consolidada por la Secretaría de Salud de la Alcaldía de Medellín, alojada en el repositorio público MEData. En general, la base de datos cuenta con información muy completa; sin embargo, sus registros datan de los años 2015 a 2018, lo cual afecta enormemente la actualidad y trazabilidad de los datos, más teniendo en cuenta que el 2019 fue el año previo a la crisis social y económica provocada por la pandemia del COVID-19, lo cual otorgaría una imagen más clara del impacto provocado por ésta en términos de nutrición infantil. La BD contiene problemas de redacción que se deben corregir y normalizar, efectuando una limpieza de las tablas.*

## 2.2 Descripción de datos (diccionario):

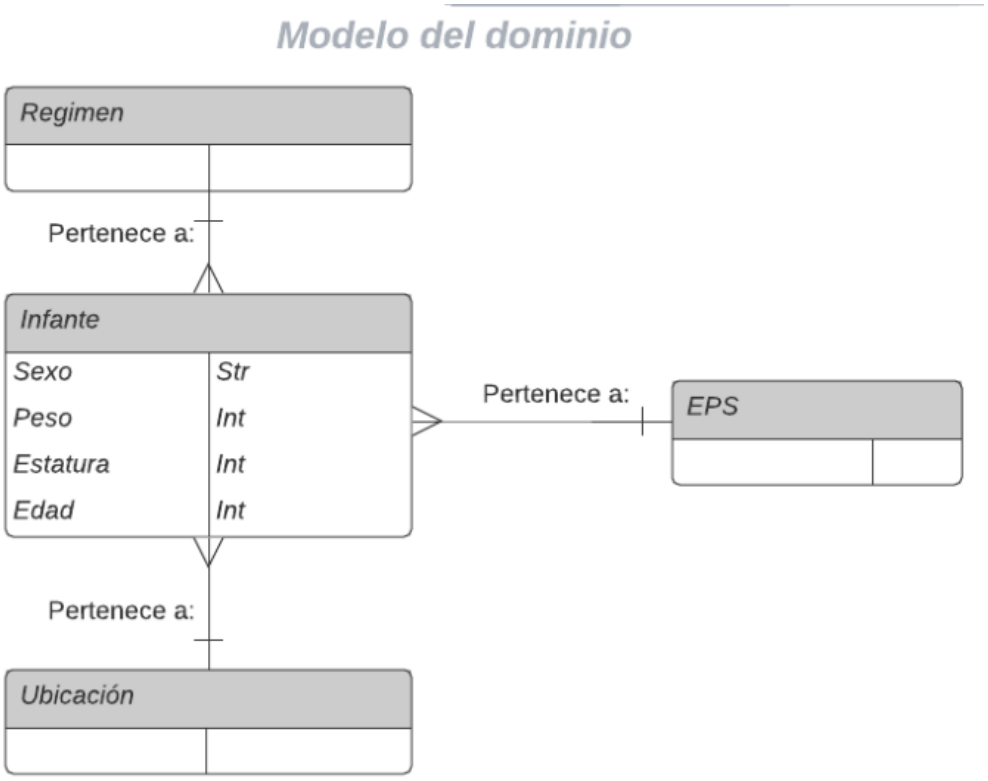
Diligencia la siguiente tabla, puede agregar otra columna si lo considera necesario.

| Nombre del atributo / variable | Formato o Tipo de Dato | Descripción |
|---|---|---|
| consecutivo | number | consecutivo único |
| régimen | string | Régimen de seguridad social |
| eps | string | Entidad promotora de salud que se clasifican de acuerdo con el régimen y puede pertenecer a uno o varios regímenes a la vez |
| sexo | string | Sexo del menor |
| peso | number | Peso en kilogramos |
| estatura | number | Estatura en centímetros |
| peso_edad_ds | number | Desviación obtenida para el indicador peso para la edad |
| peso_edad_denomina | string | Denominación del indicador peso para la edad |
| talla_edad_ds | number | Desviación obtenida para el indicador talla para la edad |
| talla_edad_denomina | string | Denominación del indicador talla para la edad |
| peso_talla_ds | number | Desviación obtenida para el indicador peso para la talla |
| peso_talla_denomina | string | Denominación del indicador peso para la talla |
| comuna | number | Comuna de residencia en Medellín |

| | | |
|---|---|---|
| zona | number | Zona corresponde a la agrupación de comunas en Medellín |
| Edad_dias | number | Edad en días |
| grupo_etario | string | Agrupación de la población de acuerdo con la edad |

## 2.3 Modelo del dominio

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

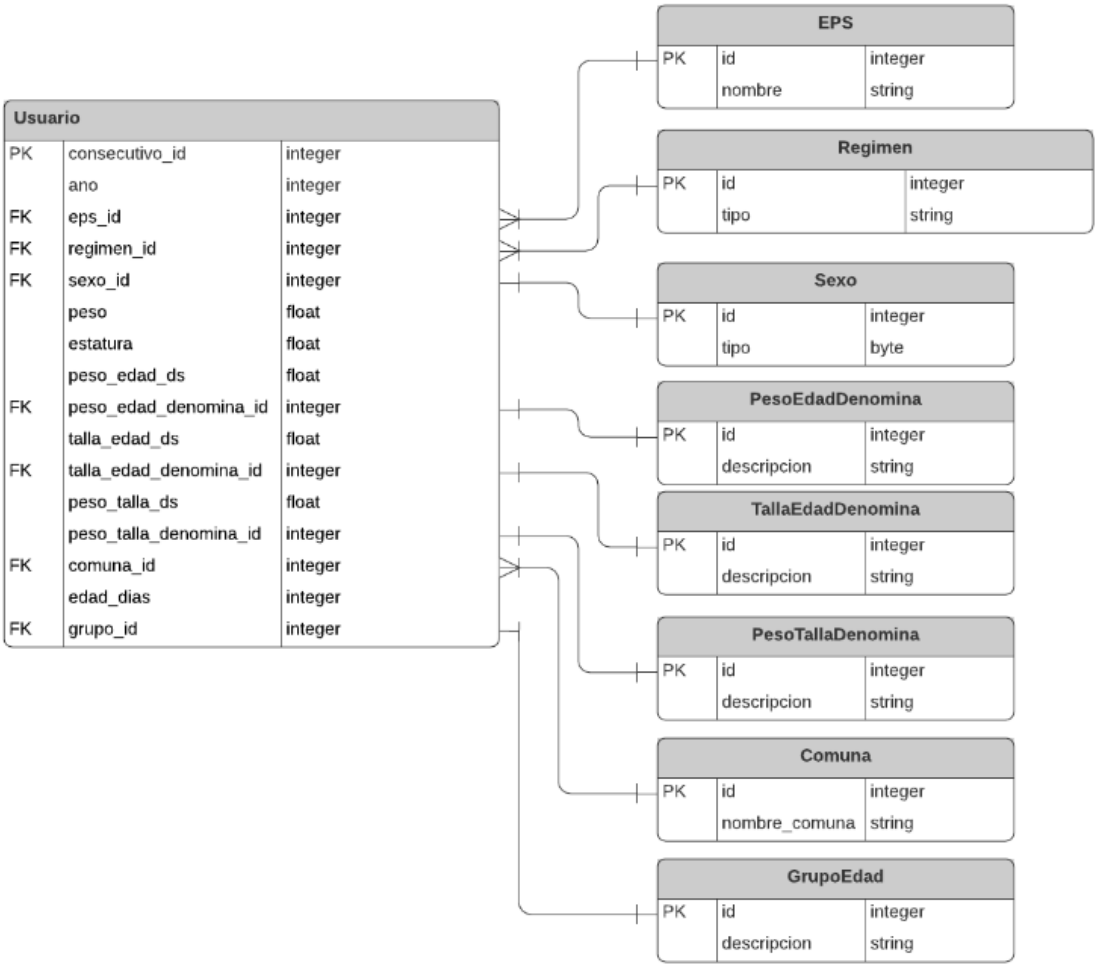



# 3 Modelo Entidad-Relación

## 3.1 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.*

Diagrama de entidad relacion de la base de datos de nutrición




## 3.2 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 creó el código para Servidor MySQL.*

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

## 3.3 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*

## 3.4 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*

# 4 MongoDB

## 4.1 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*

## 4.2 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.*

**Los pantallazos se almacenan en repositorio en el documento T1.4.2.Consultar_Datos.pdf**

# 5 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?

Cuando se toma la decisión de migrar un conjunto de datos relacionales a NoSQL, el proceso de extracción de la información original representa la primera limitación importante a considerar. Debido a que muchas de las bases de datos relacionales no están debidamente optimizadas para exportar grandes volúmenes de información en un corto periodo de tiempo, por lo que el proceso de obtención de estos datos puede llevar horas, incluso días, afectando así la productividad de la compañía. Ello también se manifiesta en un impacto en la eficiencia del sistema durante este proceso, debido a que la exportación de estos datos incluye procesos de lectura y escritura de discos físicos, comprometiendo así la eficiencia del sistema y, en consecuencia, de los procesos paralelos que se estén ejecutando.

2. ¿Cuáles limitaciones adicionales que se deben considerar, a parte de las mencionadas en el artículo?

La arquitectura misma de las bases de datos relacionales suponen una limitación para el manejo de grandes volúmenes de información, debido a que estas están altamente estructuradas y no están debidamente adecuadas para un entorno en donde la generación y almacenamiento de datos requiere de herramientas que potencien su conectividad, velocidad de consulta y eficiencia en los recursos del sistema. Por otra parte, y como es lógico, el proceso migratorio requiere de una inversión importante en términos de adaptación, capacitación y actualización de software. Por último, las bases de datos NoSQL representan un ahorro económico considerable al ser en su mayoría sistemas de código abierto (como es el caso de MongoDB), sin embargo, ello también supone una renuncia en términos de seguridad y vulnerabilidad de la información.

3. ¿Cuáles son las razones (criterios) que se deben considerar para migrar un conjunto de datos relacionados a NoSQL?

Una de las principales características de las bases de datos relacionales es que estas se constituyen en un manejo de la información a través de tablas y esquemas, en donde los registros se presentan en filas y columnas. Sin embargo, cuando se presenta un crecimiento en el volumen de información manejado, es conveniente migrar a un sistema NoSQL por dos razones principalmente: en primer lugar, las bases de datos relacionales se apoyan mucho en sus esquemas, lo que deriva en que cualquier modificación de dichos esquemas se pueda convertir en un potencial rediseño de la BD entera, o al menos en la dedicación de tiempo y recursos para su transformación; en segundo lugar, las consultas en las RDBMS implican una multiplicidad de JOINs entre tablas que contienen la información requerida, lo que afecta el proceso en términos de eficiencia, rendimiento y velocidad de consulta.

**Bibliografía**

FAO (1993). Food, Nutrition and Agriculture - International Conference on Nutrition. Roma: Food and Agriculture Organization. Consultado en: http://www.fao.org/3/U9920t/u9920t00.htm#Contents (Consultado el 15 de abril de 2021).

UN (2002). La base para el Desarrollo. Por qué los profesionales en el área del desarrollo deberían integrar la nutrición. Ginebra: United Nations.