PROFESORES:

Felipe González Casabianca

INTEGRANTES:

Maria Camila Parra Díaz (201819464)
Esteban Emmanuel Ortiz Morales (2018194646)
Sergio Julián Zona Moreno (201914936)

Tabla de contenido

1 Introducción

En el presente documento desarrollaremos el segundo proyecto del curso de Inteligencia de negocios. El propósito central es aplicar la metodología modelado dimensional para dar solución al caso de estudio presentado por “Infraestructura Visible”. Un caso enfocado en el tema de analítica y almacenamiento de defunciones en Colombia. Procederemos a desarrollar todo el proceso ETL requerido, al igual que detallar los motivos de nuestra manera de modelarlo.

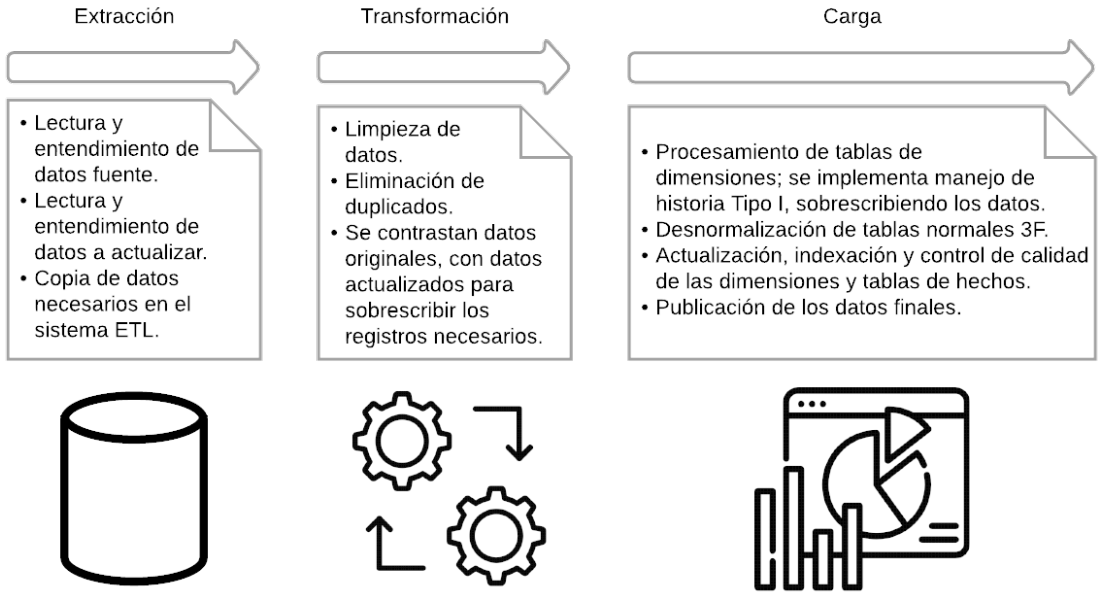
1.1 Aclaraciones preliminares

- Se recomienda al calificador/lector de este documento visualizarlo con un Zoom de 150% para evitar forzar la vista. Además, este hecho permite ver con nitidez los gráficos y las tablas presentadas.
- Adjunto con este archivo se encuentra el Notebook de operaciones realizado en Jupyter, el tablero de control de estadísticas de los datos, la presentación de resultados en formato .pdf para las organizaciones beneficiadas y el video con los resultados del proyecto.

2 Diseño y modelado

2.1 Diagrama de alto nivel del proceso ETL

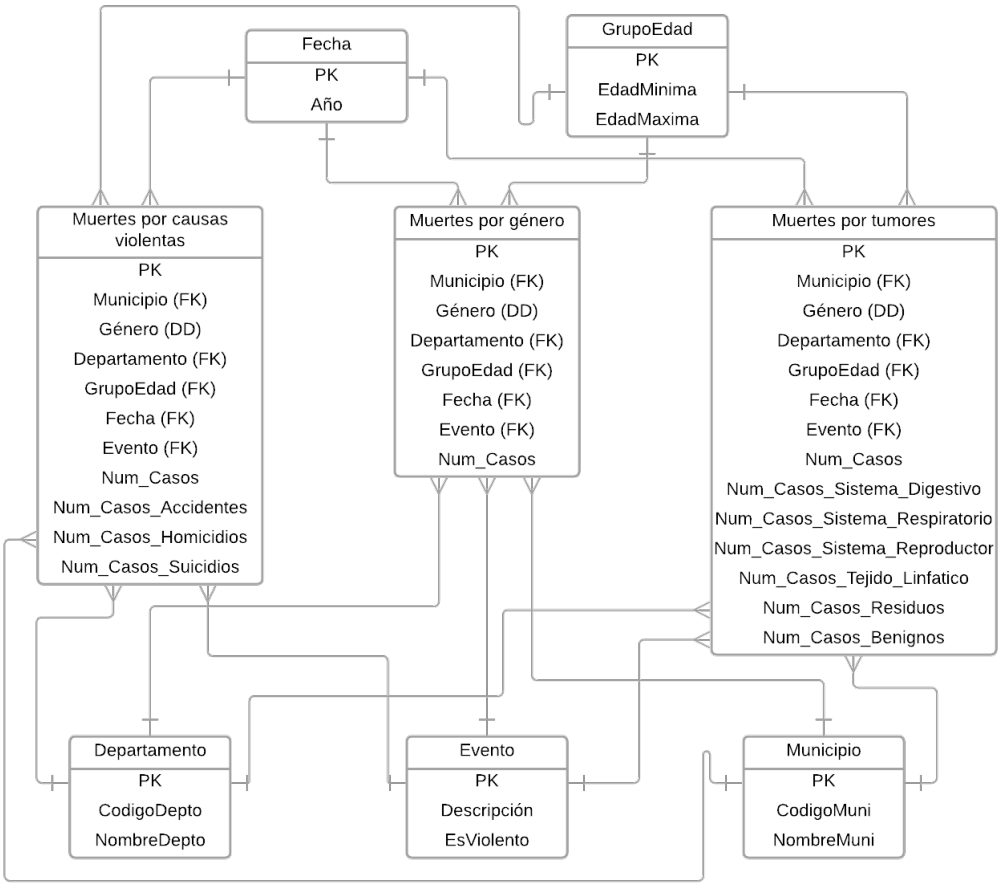
Dentro de este diagrama de alto nivel utilizamos los tres pasos de un proceso ETL. Particularmente, el modelo solamente implementa un manejo de historia Tipo I para el manejo de los atributos de las dimensiones. Durante nuestro desarrollo seguimos el flujo de trabajo del modelo planteado.



2.2 Modelado dimensional

El modelo realizado cuenta con seis dimensiones y tres tablas de hechos, que serán descritas a continuación. Cada dimensión cuenta con al menos dos atributos además de su llave primaria, a excepción de la dimensión degenerada ‘Género’. Dentro de las tablas de hechos encontramos entre una y siete medidas aditivas, las cuáles podrán

proporcionarle a negocio información valiosa relacionada con las muertes en Colombia, teniendo en cuenta las dimensiones del modelo.



2.3 Análisis en tablas de hechos

Como interés particular para negocio, optamos por modelar tres hechos fundamentales relacionados con el número de condiciones: muertes por género, muertes por desnutrición y muertes por causas violentas.

El primer hecho busca enfocarse principalmente en conocer las diferencias entre los tipos de muertes que abarcan los géneros en Colombia. Determinar cuáles son las más comunes e intentar explicar la causalidad detrás de las mismas.

El segundo hecho, busca enfocarse en conocer cuáles zonas del país se ven arduamente afectadas por problemas de desnutrición, generalmente comprendidas dentro de un marco de pobreza extrema o condiciones de difícil acceso.

Por último, el tercer hecho, busca analizar las muertes violentas dentro de un contexto polifacético, se comprende principalmente de estadísticas con muertes por homicidios, suicidios y accidentes de diversa índole. Buscamos encontrar una distribución porcentual demográfica que indique cuáles zonas del país se ven más embebidas en dichos tipos de eventos (como es lógico, normalizando los datos para efectuar un análisis adecuado entre ciudades y municipios).

2.3.1 Granularidad y justificación

Muertes por género:

Hecho con baja granularidad, si bien realiza una pequeña subdivisión según el género del individuo fallecido, toma en cuenta todos los tipos de posibles eventos de muertes, lo que abarca una gran cantidad de datos. Aunque parezca un análisis poco útil, realmente se busca encontrar cuáles son los tipos de muertes que diferencia a los grupos y determinar si existe una influencia fenotípica o genotípica en los mismos; o, por el contrario, si existe una influencia determinada por condiciones geográficas y/o demográficas.

Muertes por tumores:

Hecho con alta granularidad, especifica particularmente eventos muy puntuales, los cuáles son las muertes por tumores. Aquí se divide en 5 grupos: muertes por tumores en sistema digestivo, en sistema respiratorio, en sistema reproductor, en tejido linfático, residuos de tumores y tumores benignos. Como nos enfocamos particularmente en este nicho, se pueden obtener análisis muy interesantes por determinados grupos de edades y ubicaciones geográficas.

Muertes por causas violentas:

Hecho con medio granularidad, toma diversos tipos de causas de muerte y no se centra con claridad en uno solo. Sin embargo, este análisis facilitar la búsqueda de epicentros puntuales donde pueden ocurrir con más frecuencia estos tipos de muertes, con el fin de prevenirlas a futuro.

2.3.2 Medidas y hechos

Muertes por género (Num_Casos):

Esta variable mide la cantidad de muertes que ocurren dependiendo del género (hombre – mujer). Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por tumores (Num_Casos):

Esta variable mide la cantidad de muertes que ocurren debido a la desnutrición. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por tumores - sistema digestivo (Num_Casos_Sistema_Digestivo):

Esta variable mide la cantidad de muertes de personas por tumores en el sistema digestivo. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por tumores - sistema resp. (Num_Casos_Sistema_Respiratorio):

Esta variable mide la cantidad de muertes de personas por tumores en el sistema respiratorio. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por tumores - sistema repr. (Num_Casos_Sistema_Reproductor):

Esta variable mide la cantidad de muertes de personas por tumores en el sistema reproductor. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por tumores - tejido linfático (Num_Casos_Tejido_Linfático):

Esta variable mide la cantidad de muertes de personas por tumores en el tejido linfático. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por tumores - residuos (Num_Casos_Residuos):

Esta variable mide la cantidad de muertes de personas por residuos de tumores. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por tumores - benignos (Num_Casos_Benignos):

Esta variable mide la cantidad de muertes de personas por tumores benignos. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por causas violentas (Num_Casos):

Esta variable mide la cantidad de muertes que ocurren por causas violentas. Por causas violentas se entienden: accidentes en transporte terrestre, caídas, accidentes causados por máquinas, exposición al humo, fuego y llamas, etc. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por causas violentas - accidentes (Num_Casos_Accidentes):

Esta variable mide la cantidad de muertes que ocurren por causas violentas relacionadas con accidentes. Por ejemplo: accidentes en transporte terrestre, caídas, accidentes causados por máquinas, ahogamiento y sumersión accidentales, etc. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por causas violentas - homicidios (Num_Casos_Homicidios):

Esta variable mide la cantidad de muertes que ocurren por causas violentas relacionadas con homicidios (agresiones, inclusive secuelas). Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por causas violentas - suicidios (Num_Casos_Suicidios):

Esta variable mide la cantidad de muertes que ocurren por causas violentas relacionadas con suicidios (lesiones autoinfligidas intencionales, inclusive secuelas). Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

2.4 Análisis de dimensiones y atributos

Dimensión fecha:

La tabla de fechas es importante porque mediante esta se podrá hacer referencia a información puntual dentro del modelo. Gracias a esto se podrán realizar búsquedas sobre fechas especificas o rangos de fechas y hacer análisis más profundos sobre los resultados de búsqueda.

La fuente de datos principal cuenta con una columna “año” de la cual se nutrirá la tabla de fechas, por lo que el único atributo con el que contará dicha tabla será el año.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión departamento:

Dentro de las necesidades de análisis en el proceso ETL, el hecho de poder hacer un análisis sobre los datos de muertes basados en la ubicación geográfica es trascendental dentro de los objetivos del negocio. En este caso la diferenciación geográfica basada en departamento ayuda a establecer correlaciones sólidas entre el departamento y los tipos de muerte.

Dentro del CSV utilizado para el análisis existe una columna llamada departamento que contiene el nombre del departamento asociado a un registro especifico dentro de los datos, de esta columna justamente se extraerá el atributo nombre que constituye a el único atributo dentro de la dimensión además de su llave privada.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión municipio:

Como se aclaró previamente en la dimensión departamento, poder establecer correlaciones entre hechos y ubicaciones geográficas da como resultado información muy valiosa para el negocio. Puesto que dicha información y los respectivos análisis que se pueden hacer con base en ella son muy atractivos para instituciones estatales.

Los datos de muertes que fueron otorgados para este análisis cuentan con una columna llamada municipio y dentro de ella se especifica el nombre el municipio. A partir de esta columna se estableció el atributo único de la tabla de dimensión municipio, su nombre.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión evento:

Probablemente la dimensión más importante dentro de los requerimientos analíticos del negocio. Puesto que el análisis de los diferentes tipos de muertes en el país es la base sobre la que se funda el estudio que se quiere proponer. A partir de la diferenciación de eventos de muerte es que toman sentido las demás dimensiones y se pueden hacer búsquedas más complejas, completas y provechosas que resolverán los problemas planteados por el negocio.

En función de la tabla de datos suministrada, se extrajo el contenido de la columna evento y se convirtió en el atributo descripción de su dimensión. Adicionalmente se consideró importante agregar otro atributo, uno de tipo booleano llamado EsViolento que permita hacer una distinción dentro de los eventos y posibilite hacer análisis sobre las muertes violentas, uno de los objetivos principales planteados en el negocio. Dicho atributo se asignará arbitrariamente a todo evento que tenga una descripción que no se asocie a una muerte relacionada a causas naturales.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión género:

Esta dimensión es de gran importancia para negocio para poder realizar análisis relacionados con el género. Esta información es bastante relevante para el contexto actual del país, puesto que, por ejemplo, se cuenta con un exceso de muertes por violencia de género en mujeres. Tener la posibilidad de realizar análisis acerca de estas temáticas permiten tener un mejor acercamiento a estas problemáticas para entenderlas de una forma más efectiva y clara.

Dentro de los atributos encontramos únicamente su llave primaria y el tipo, donde se indica si el género es masculino o femenino.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión grupo edad:

Con esta dimensión negocio puede encontrar información relevante de acuerdo a la edad de las víctimas. Se podrán encontrar medidas que brinden un panorama más claro de cómo se ven afectadas las diferentes poblaciones de Colombia de acuerdo con su edad. De esta forma se podrá saber cuál es el tipo de incidente que afecta en mayor medida a las personas de la tercera edad, o cuál es la cantidad de menores de edad que mueren por aflicciones como la desnutrición.

Para esta dimensión encontramos los atributos de edad mínima y edad máxima, que nos permitirán delimitar los rangos para los grupos de edades para posterior procesamiento de los datos y análisis de los mismos.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

3 Perfilamiento de datos, creación de la base de datos y procesos ETL

Para nuestra implementación utilizamos dos herramientas: Python para preprocesamiento y organización de las dimensiones. Y BigQuery para subir la información a una base de datos y poder almacenar nuestro modelo final. A continuación, se detallan los pasos:

3.1 Perfilamiento de datos

Se presentaron las siguientes modificaciones en los datos:

- Dado que nuestras tablas de hechos no comprenden análisis que requieran de los totales (si no que se pueden obtener por de medio de Querys simples), se opta por eliminar los registros que contienen dicha información.
- Se ajustan los códigos en la columna de departamentos para unificar ciertos valores que se encontraban con distintos formatos.
- Ajuste de los rangos de edades para la dimensión de GrupoEdad.
- Creación del atributo EsViolento para la dimensión de eventos.
- Creación de métricas aditivas en las columnas de hechos.

Al finalizar el proceso, generamos dimensiones no normalizadas, con métricas redundantes pero adecuadas para el proceso ETL y, tablas de hechos que agrupan ciertos perfiles de muertes en Colombia que son de interés para el negocio. Esto puede ser visualizado en el cuadernillo adjunto en el repositorio de anexos.

3.2 Creación de la base de datos y carga de datos

Después de exportar el modelo perfilado del cuadernillo en Jupyter Lab, subimos nuestra información la plataforma de Google BigQuery, que cuenta con una base de datos integrada. Allí importamos las tablas las cuales crearon los Constraints necesarios entre llaves foráneas y generaron una primera instancia del modelo. En el repositorio adjunto, en la carpeta ‘Proceso ETL’ se encuentran capturas del desarrollo en BigQuery.

3.3 Implementación de los procesos ETL

3.3.1 Herramienta seleccionada y justificación

Para el desarrollo del proceso ETL se decidió trabajar con Google Cloud BigQuery debido a que se utiliza el dialecto SQL estándar, con el cual los integrantes del grupo nos encontramos bastante familiarizados. Adicionalmente, aunque la curva de aprendizaje es de alta complejidad, es más sencillo que aprender a utilizar herramientas como Spoon y Talend. Finalmente, debido a que es una herramienta serverless, BigQuery permite manejar grandes cantidades de datos, por lo que es ideal para las fuentes de datos a utilizar en el proyecto.

3.3.2 Diseño y transformaciones

La mayoría de las transformaciones de desnormalización fueron efectuadas en la herramienta de Python (véase el cuadernillo). Sin embargo, el manejo de historia Tipo I (sobre-escritura de los datos), fue implementado específicamente en la herramienta de Google BigQuery. Adicionalmente, en Google BigQuery también efectuamos un Join entre cada tabla de hechos y sus dimensiones para crear tres *tablas maestras* finales, que contienen toda la información necesaria para efectuar Queries avanzadas. Estas tablas son particularmente útiles porque contienen toda la información de nuestro Dataware House. En el repositorio adjunto, dentro de la carpeta 'Tablas/Tablas Maestras' se pueden visualizar los .csv generados.

4 Arquitectura de solución

4.1 Herramienta seleccionada y justificación

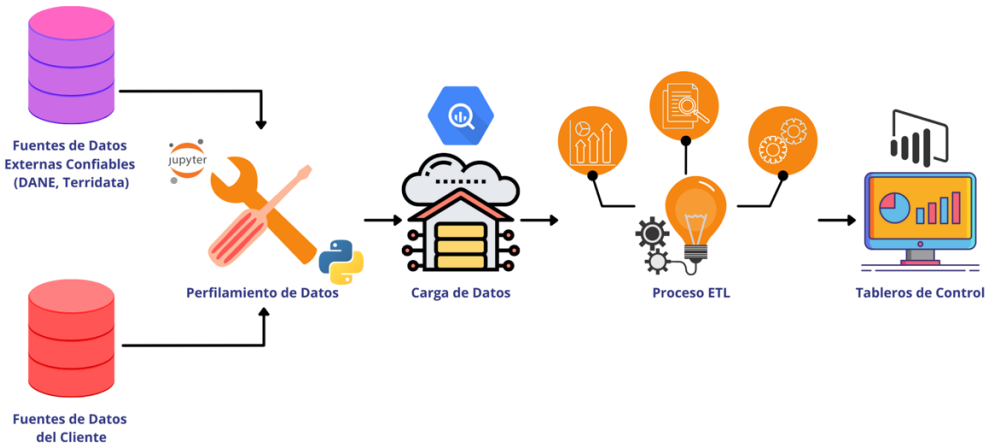
Para el desarrollo de la arquitectura de solución se decidió trabajar con Power BI por varias razones. Por ejemplo, Power BI cuenta con un diseño mucho más flexible que otras herramientas, como Tableau, que permite al usuario operar con grandes cantidades de datos de manera eficiente. Además, Power BI cuenta con la capacidad de integrarse con cualquier herramienta de Microsoft, como Excel, lo que resulta útil para el manejo de los datos. La aplicación cuenta con una galería de gráficos que permite presentar una visualización atractiva en los tableros de control para negocio.

4.2 Diseño los tableros de control

Optamos por diseñar un tablero de control para cada una de las tablas de hechos, con el fin de sintetizar información que consideramos relevante para el negocio. Los hallazgos relevantes se encuentran dentro de la presentación, a la cual se puede acceder a través del enlace que se encuentra en los anexos. Adicionalmente, en el repositorio adjunto, también en los anexos, se pueden evidenciar estos tres tableros.

4.3 Propuesta de arquitectura de solución

Para la arquitectura de solución, se decidió generar una propuesta que tuviera en cuenta cada uno de los procesos realizados en la primera etapa del proyecto: acceso a las fuentes de datos, perfilamiento, carga, proceso ETL y tableros de control. Se propone, entonces, una herramienta específica para cada tarea del proceso, incluyendo Google BigQuery, JupyterNotebook y Power BI.



5 Repartición de tareas y puntos asignados por integrante

Integrantes	Tareas realizadas	Número de horas	Puntos
María Camila Parra Díaz (201819464)	• Diseño del tablero de control de la tabla de hechos “Muertes por género”. • Diseño de la presentación para negocio. • Redacción del documento. • Diseño e implementación del modelo ETL, del modelo dimensional y de la arquitectura de solución.	25	100/3
Esteban Emmanuel Ortiz Morales (201913613)	• Diseño del tablero de control de la tabla de hechos “Muertes por causas violentas”. • Diseño de la presentación para negocio. • Prueba de calidad y revisión del documento. • Diseño e implementación del modelo ETL, del modelo dimensional y de la arquitectura de solución.	25	100/3
Sergio Julián Zona Moreno (201914936)	• Diseño del tablero de control de la tabla de hechos “Muertes por tumores”. • Vídeo de presentación a negocio. • Consolidación del documento. • Diseño e implementación del modelo ETL, del modelo dimensional y de la arquitectura de solución.	25	100/3

Todos trabajaron de manera adecuada y realizaron aportes significativos e importantes en el proyecto. Por dicho motivo, se dividen los puntos de manera equitativa para todos los integrantes.

6 Bibliografía

- Universidad de los Andes. (2020). *Material del curso: Inteligencia de negocios*. Disponible en BLOQUE NEÓN para estudiantes de la Universidad.

7 Anexos

Enlace al repositorio:

https://github.com/SergioZona/Proyecto2_BI_ee_ortiz_mc_parrad_sj_zona

Enlace de vídeo de presentación:

https://youtu.be/s7qvIGYSqVI

Enlace de la presentación:

https://www.canva.com/design/DAEwJOwmpnc/w_ZvIbDURrQH7kcEND9JMg/view?utm_content=DAEwJOwmpnc&utm_campaign=designshare&utm_medium=link&utm_source=sharebutton