Santiago de Cali, 11 de Octubre de 2023.

Señores
Pontificia Universidad Javeriana Cali.
Ph.D. Luisa Rincón
Directora Maestría en Ingeniería de Software.
Cali.

Cordial Saludo.

Por medio de la presente hago constar que en mi calidad de director de trabajo de grado he revisado el proyecto titulado “Herramienta para el análisis de los metadatos del portal Open Data Colombia” realizado por el estudiante de Maestría en Ingeniería de Sistemas y Computación Diego Fernando Segura Herrera (cod: 8972927), el cual se encuentra terminado y considero que cumple con los requisitos para ser sustentado.
Atentamente,

Ph.D Christian Arias

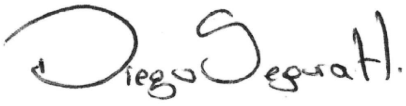
Santiago de Cali, 11 de Octubre de 2023.

Señores
Pontificia Universidad Javeriana Cali.
Ph.D. Luisa Rincón
Directora Maestría en Ingeniería de Software.
Cali.

Cordial Saludo.

Me permito presentar a su consideración el proyecto de grado titulado “Herramienta para el análisis de los metadatos del portal Open Data Colombia” con el fin de cumplir con los requisitos exigidos por la Universidad y para que sea sometido a revisión del jurado y cumpla su aprobación, para conseguir posteriormente el título de Master en Ingeniería de Software.

Atentamente,



Diego Fernando Segura Herrera
Código: 8972927

Maestría en Ingeniería de Software
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: "Herramienta para el análisis de los metadatos del portal Open Data Colombia"

1. ÉNFASIS: Ingeniería de Software
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Análisis de Datos
4. ESTUDIANTE (S): Diego Fernando Segura Herrera
5. CORREO ELECTRÓNICO: dsegura@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Calle 10ª # 78ª-09 Cel: 3186552774
7. DIRECTOR: Christian Arias
8. VINCULACIÓN DEL DIRECTOR: Cátedra
9. CORREO ELECTRÓNICO DEL DIRECTOR: christian.arias@javerianacali.edu.co
10. CO-DIRECTOR(ES) (Si aplica): N/A
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica): N/A
12. OTROS GRUPOS O EMPRESAS: N/A
13. PALABRAS CLAVE (al menos 5): Metadatos, Open Data, Datos Abiertos, Business Intelligence, ETL, Percepción de Utilidad.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Educación de Calidad, Industria, Innovación e infraestructura
15. FECHA DE INICIO (Desarrollo del proyecto): 13/10/2023
16. RESUMEN (máximo 400 palabras).

Los datos abiertos tienen un papel crucial en el contexto de la transparencia de las entidades, y más aún cuando son Gubernamentales. Por esta razón fueron creadas las plataformas de datos abiertos en diferentes Gobiernos a nivel global. Particularmente en Colombia se creó la plataforma datos.gov.co en el año 2016 para ser usada como la plataforma oficial de datos abiertos.

Estas plataformas deben cumplir con ciertos lineamientos y propósitos. Generar conocimiento e información que ayude al crecimiento no solo del Gobierno si no de los particulares e investigadores que usan estos datos en diferentes contextos. La utilidad de estos datos está estrechamente ligada a la calidad de la plataforma, ya que debe proveer mecanismos para validar los datos que se ingresan en ella. Es por esta razón que se pretende crear una herramienta que brinde indicadores para evaluar la utilidad de los datos de la plataforma.

Para ello se plantea construir un modelo dimensional, que permita generar métricas sobre los metadatos existentes dentro de la plataforma datos.gov.co y buscar relaciones entre los valores de los metadatos y la percepción de utilidad de los usuarios de la plataforma.

Pontificia Universidad Javeriana Cali
Facultad de Ingeniería.
Ingeniería de Sistemas y Computación.
Proyecto de Grado.

Herramienta para el análisis de los metadatos del portal Open Data Colombia

Diego Fernando Segura Herrera

Director: Ph.D Christian Arias

11 de Octubre de 2023

Índice general

Índice de figuras

Índice de tablas

0.1. Resumen

Los datos abiertos tienen un papel crucial en el contexto de la transparencia de las entidades, y más aún cuando son Gubernamentales. Por esta razón fueron creadas las plataformas de datos abiertos en diferentes Gobiernos a nivel global. Particularmente en Colombia se creó la plataforma *datos.gov.co* en el año 2016 para ser usada como la plataforma oficial de datos abiertos.

Estas plataformas deben cumplir con ciertos lineamientos y propósitos. Generar conocimiento e información que ayude al crecimiento no solo del Gobierno si no de los particulares e investigadores que usan estos datos en diferentes contextos. La utilidad de estos datos está estrechamente ligada a la calidad de la plataforma, ya que debe proveer mecanismos para validar los datos que se ingresan en ella. Es por esta razón que se pretende crear una herramienta que brinde indicadores para evaluar la utilidad de los datos de la plataforma.

Para ello se plantea construir un modelo dimensional, que permita generar métricas sobre los metadatos existentes dentro de la plataforma *datos.gov.co* y buscar relaciones entre los valores de los metadatos y la percepción de utilidad de los usuarios de la plataforma.

Palabras Clave: Metadatos, *Open Data*, Datos Abiertos, *Business Intelligence*, ETL, Percepción de Utilidad.

0.2. Abstract

Open data has a crucial role in the context of the transparency of the entities, and even more so when they are Governmental. For this reason, open data platforms were created in different governments at the world level. Particularly in Colombia, the platform datos.gov.co is created in 2016 to be used as the official open data platform.

These platforms must comply with certain guidelines and purposes. Generate knowledge and information that helps the growth not only of the Government but not of individuals and researchers who use these data in different contexts. The usefulness of these data is strictly linked to the quality of the platform since it must provide mechanisms to validate the data that they enter it. It is for this reason that it is intended to create a tool that Provides indicators to assess the usefulness of the platform data.

To do this, it is proposed to build a dimensional model that allows generating metrics on the existing metadata within thedata.gov.co platform and searching for relationships between the values of metadata and the perception of usefulness of platform users.

Keywords: Metadata, *Open Data*, *Business Intelligence*, ETL, Perceived Usefulness.

CAPÍTULO 1

Introducción

1.1. Definición del problema

En la actualidad los datos han tomado un papel fundamental en casi todos los aspectos empresariales y gubernamentales para la toma de decisiones. También dentro de la comunidad científica se ha convertido en una de las especialidades de las ciencias de la computación más relevantes en los últimos años en el mundo entero. Dado este escenario se ha generado un movimiento que busca que los datos sean compartidos de manera abierta, para que las personas e instituciones puedan hacer uso de esos datos con diferentes propósitos, que van desde la generación de nuevo conocimiento visto desde los ámbitos académicos, hasta predicciones para comportamiento de mercado en entornos empresariales.

Este movimiento es el de Open Data, esta expresión de Open Data como concepto, no es nueva, pero su definición formal sí lo es. Según Murray-Rust (2008) es un término usado por la comunidad científica para los datos que pueden ser publicados y usados sin tener barreras o limitaciones. Open Data ha sido adoptado por diferentes entidades alrededor del mundo en diferentes iniciativas de carácter público y privado.

En el caso particular de Colombia existen diferentes iniciativas gracias a la ley 1792 de 2014, que obliga a todas las entidades públicas a divulgar sus datos, adoptando así los principios que establece la Carta Internacional de Datos Abiertos (Charter, 2015). El Gobierno Colombiano en su programa Gobierno Digital ejecutado a través del Ministerio de las Tecnologías de la Información y las Comunicaciones, promueve y habilita las condiciones para el uso y generación de datos abiertos del estado colombiano en la plataforma Datos Abiertos (Gobierno de Colombia, 2022b).

A través de la plataforma Datos Abiertos, el gobierno colombiano dispone los datos de diferentes sectores para que estos sean usados y expuestos. Estos datos al ser de diferentes fuentes y tópicos, dependiendo del sector, del negocio o de la entidad territorial que los comparta, podrían tener una importante cantidad de usos, y si bien la plataforma ofrece opciones de consultar los datos con filtros básicos como año, región, rango de fechas, entre otros (Gobierno de Colombia, 2022a), no ofrece alternativas para evaluar los datos a partir de los metadatos dispuestos.

Contar con este tipo de evaluaciones es importante para poder mejorar la iniciativa y poder cumplir de mejor manera con los objetivos de Open Data en Colombia. Ya que, también hay que tener en cuenta que tampoco es claro si la utilidad de estos datos es buena o es suficiente para que genere valor a los actores que usan la plataforma.

1.2. Planteamiento del problema

- ¿Cómo medir la utilidad de los datos de la iniciativa Datos Abiertos en Colombia a partir de sus metadatos?.
- ¿Cómo evaluar los metadatos del portal de Datos Abiertos del gobierno colombiano?.

1.3. Objetivos

1.3.1. Objetivo General

Crear una herramienta de software que brinde indicadores para evaluar la utilidad de los datos disponibles en la plataforma *datos abiertos* del Gobierno Colombiano a partir de sus metadatos.

1.3.2. Objetivos Específicos

- Entender los metadatos disponibles de la plataforma *datos abiertos* y plantear indicadores para el proceso de evaluación
- Diseñar y construir un modelo dimensional para almacenar los metadatos relevantes para el análisis.
- Implementar un proceso de extracción de los metadatos de la plataforma *datos abiertos* para almacenarlos en el modelo dimensional.
- Construir visualizaciones de los metadatos para evaluar la utilidad de los datos disponibles.
- Evaluar la percepción de utilidad de los datos de la plataforma *datos abiertos* a partir de los metadatos obtenidos por la herramienta propuesta.

CAPÍTULO 2

Marco de referencia

2.1. Bases teóricas

Open Data (Datos Abiertos)

El concepto de Open Data está fuertemente arraigado en la filosofía de compartir los recursos para diferentes fines, indicando que estos recursos se pueden usar, reusar y distribuir sin tener limitaciones de derechos de autor o algún tipo de consecuencia legal, a lo sumo se debe indicar de quién son Open Data HandBook (2022). También hace parte de ese concepto que los datos sean brutos “sin procesar"para que así cualquier persona pueda sacar análisis o conclusiones sin sesgos de ningún tipo.

Desde los años 2000 se viene dando definición a este concepto, según Kassen (2013) lo cataloga como un fenómeno que se da a partir de las iniciativas del gobierno de los Estados Unidos de desclasificar y compartir información al público. Otros desde el punto de vista científico y el uso de los datos en este ámbito, se familiarizan más con la definición de compartir sin limitaciones como es el caso de Murray-Rust (2008).

Sin embargo, se han creado instituciones que ofrecen una definición y guían el entendimiento de Open Data, brindando un contexto más profundo al respecto, es el caso de Open Knoledge Fundation (2022) y Open Data HandBook (2022), que proporcionan definiciones más explicitas sobre cuando en realidad los datos son “Open”y que condiciones deben cumplir para ser usados correctamente en los entornos donde se utiliza esta terminología, desde ambientes empresariales hasta académicos.

Business Intelligence (BI)

La inteligencia de negocio es un proceso que ayuda al análisis de las condiciones actuales de una compañía que ayudan en la reducción de costos, mejoramiento de procesos y calidad de servicio entre otros(Foley and Guillemette, 2010). Qué está pasando y por qué, basándose en el uso de los datos generados por la compañía, usualmente con herramientas de bodegas de datos y modelos dimensionales para estructurar la información que se genera.

Modelo Dimensional

Un modelo dimensional es un modelo de datos que amplia y correlaciona diferentes aspectos de los procesos de un negocio. Este tipo de modelos está especialmente hecho para responder a demandas de grandes volúmenes de datos y están basados en esquemas dimensionales como estrella y copo de nieve (IBM, 2022).

El esquema en estrella es el bloque fundamental para construir modelos dimensionales, este esquema consiste en una tabla central, llamada tabla de hechos y una serie de tablas mas pequeñas al rededor de la tabla central. Estas tablas son llamadas tablas de dimensiones.

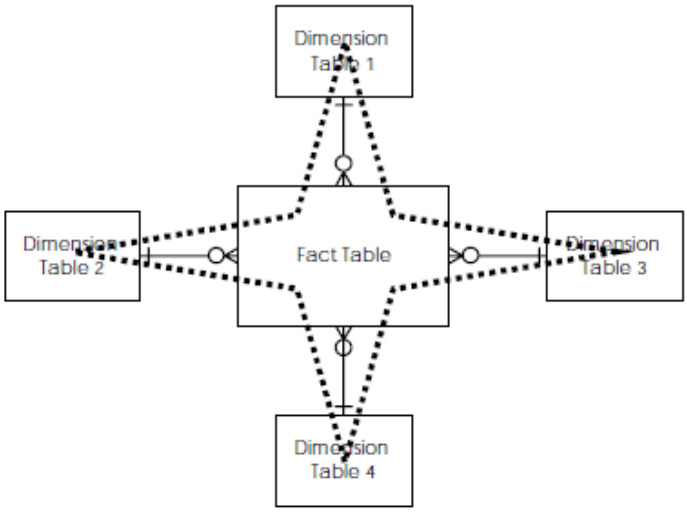


Figura 2.1: Esquema en estrella. Fuente: Moody et al. (2000)

Como se observa en la Figura 2.3 la tabla de hechos esta en el centro y las tablas de dimensiones estan a su alrededor mostrandose como una estrella, por esta raon el nombre del esquema.

Por otra parte el esquema copo de nieve, añade un poco mas de complegidad desde el punto de vista que muestra más jerarquias del modelo y en palabras de Ralph Kimball (R.K, 1996) no es deseable ya que añade complejidad a los queries y al esquema y va en contra de producir un diseño simple y amigable para los usuarios. La Figura 2.2 ilustra un esquema de copo de nieve.

Estos modelos han sido usados en el ámbito empresarial para generar valor evaluando grandes cantidades

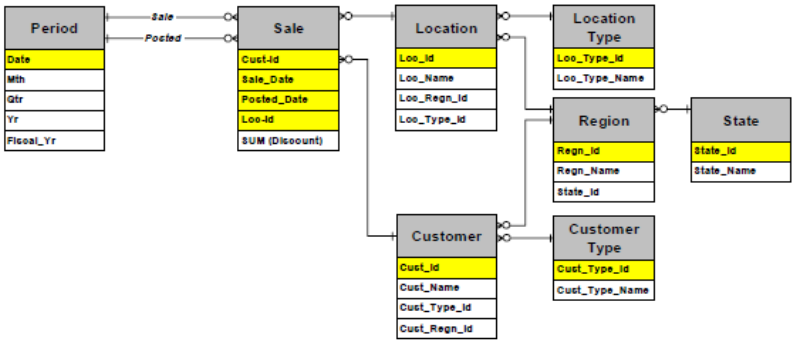


Figura 2.2: Esquema en copo de nieve. Fuente: Moody et al. (2000)

de datos en diferentes, ayudando a las empresas a tomar decisiones estratégicas mediante procesos de Business Intelligence usando comúnmente Bodegas de Datos que son colecciones de volúmenes importantes de información y Data Marts que son pequeñas bodegas de datos especializadas por temas específicos.

Extracción, Transformación y Carga

Es el proceso a través del cual los datos son leídos, moldeados, transformados y cargados de una fuente (Base de datos, Archivos, Flujos de datos) a otra. Más conocido por sus siglas en ingles ETL(Extract, Transform and Load). Muy utilizado por las organizaciones en sus procesos de Business Intelligence en el análisis de sus datos para la toma de decisiones estratégicas de negocio, también es un proceso utilizado para apoyar la operación de sistemas de información, utilizando las ETLs como procesos integradores entre sistemas heterogéneos. Cambiando los tipos de datos, uniendo información en diferentes formatos y unificando la información que se requiera para cualquier proceso relacionado con integrar datos.

Percepción de utilidad

Es una medida propuesta para añadir un valor más objetivo a la aceptación de uso de un producto de software, introduciendo esta métrica se propone hacer menos subjetivo el uso de una herramienta de software (Davis, 1989).

Para realizar la medición se propone la creación de un instrumento que permita recolectar información sobre el uso de un producto de software determinado, este instrumento en este caso es un cuestionario con preguntas que van orientadas a los usuarios quienes proveen esta información para que sea analizada con el fin de determinar si es útil o no o si de alguna manera ayuda al usuario a mejorar las actividades relacionadas con su trabajo o investigación.

Estos mecanismos de medición se han venido usando como apoyo para escoger software que acelere o

mejore el rendimiento de los procesos de las personas dentro de su trabajo (Council, 1985).

2.2. Estado del arte

Datos Abiertos Colombia

En el entorno de la República de Colombia y como consecuencia de la tendencia a nivel global respecto del uso de los datos y de la transparencia de los gobiernos bajo el uso de la figura de publicar los datos para que sean de dominio público.

Se han presentado iniciativas(plataformas) desde el gobierno nacional a partir de la Ley 1792 de 2014 (Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional) (Gobierno de Colombia, 2014) bajo los principios de transparencia, buena fe, facilitación, no discriminación y gratuidad, entre otros y define el marco para la disposición de la información a través de medios digitales a la ciudadanía. También se definen los datos abiertos como "*todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos*".

En este contexto la iniciativa open data Colombia (www.datos.gov.co) es creada en el año 2016 en cabeza del Ministerio de las Tecnologías y las Telecomunicaciones con el fin de introducir a Colombia dentro de la comunidad de países que cuentan con una plataforma de datos abiertos.

Otras plataformas de datos abiertos a nivel departamental y local existen en Colombia, pero en la que se enfocará este documento es en la iniciativa nacional datos.gov.co.

2.2.1. Antecedentes

Con respecto a la calidad de los datos y de las plataformas de datos abiertos, existen diferentes metodologías que proponen mecanismos para evaluar ambos aspectos. La metodología de cinco estrellas propuesta por Tim Berners-Lee la cual evalúa el portal desde la accesibilidad y reusabilidad usando 5 niveles según la cantidad de estrellas (Tim Berners-Lee, 2006). La Figura 2.3 ilustra cada nivel.

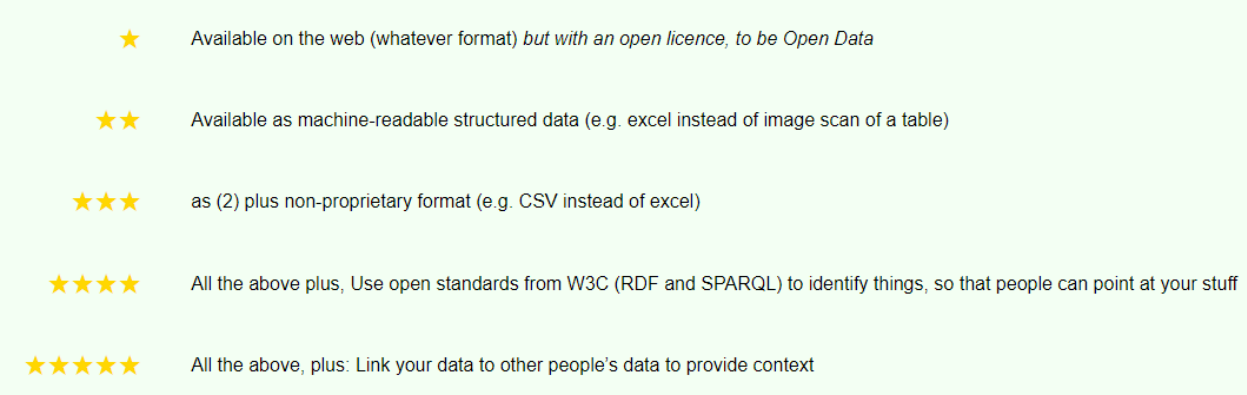


Figura 2.3: Modelo de cinco estrellas Tim Berners-Lee.
Fuente: Tim-Berners-Lee (2022)

Otra metodología para evaluar reutilización de datos abiertos es la propuesta Meloda (Abella et al., 2014) que propone hacer la evaluación desde cuatro dimensiones, estándares técnicos, acceso, legal, y modelo de datos, La Figura 2.4 muestra cada una de las cuatro dimensiones.

Estándares técnicos	Acceso	Legal	Modelo de datos
1. Estándar privativo Ej.: .xls, .shp, .doc	1. Sin acceso Ej.: mail no automático o acceso en persona	1. *Copyright* Ej.: *copyright*	1. Sin modelo publicado Ej.: tabla de datos sin descripción de los campos
2. Estándar abierto Ej.: .csv, .ods, wms	2. Acceso vía web con registro Ej.: formulario manual	2. Uso privado Ej.: *copyright* permitiendo uso personal	2. Modelo con campos de datos Ej.: tabla de datos con descripción de los campos
3. Estándar abierto con metadatos Ej.: rdf, rss, json	3. Acceso directo vía web Ej.: url único	3. Uso no comercial Ej.: CC BY-NC 4.0	3. Modelo con especificaciones de campos Ej.: vocabularios disponibles
	4. Acceso vía web con parámetros Ej.: url con parámetros	4. Uso comercial Ej.: CC BY-SA 4.0	4. Modelo externo normalizado Ej.: vocabularios disponibles aceptados por organización de normalización
	5. Acceso completo (API) Ej.: punto de acceso *Sparql*	5. Uso no limitado con autoría Ej.: CC BY 4.0	5. Modelo externo y generalizado Ej.: vocabularios disponibles aceptados por organización de normalización y reconocidos

Figura 2.4: Dimensiones modelo Meloda. Fuente: Abella et al. (2014)

Algunos autores han propuesto metodologías para evaluar la plataforma *Datos Abiertos* de Colombia, creando mediciones basadas en metodologías como la Meloda y la de cinco estrellas, para aplicarlas y determinar la calidad de la plataforma y de los datos, recolectando la información de grupos de usuarios.

Un ejemplo de esto es el artículo "*Proposal for the Evaluation of Open Data Portals*"(Melo and Sanabria, 2020), donde se propone un mecanismo de medición basado en metodologias conocidas como la de cinco estrellas y la Meloda, aplicándolas al caso de estudio de la plataforma *Datos Abiertos*.

Sin embargo, para el contexto específico de la plataforma *Datos Abiertos* aún es incierto que tan útil es la información dispuesta en la plataforma. Ya que se han hecho algunas aplicaciones de metodologías, pero no se puede concluir si la percepción de utilidad de la plataforma y de los datos es favorable o por el contrario es deficiente.

En los siguientes capítulos se mostrara el proceso de elaboración y medición que se aplicará a la plataforma *Datos Abiertos*, a través de la elaboración de un modelo dimensional que será usado para almacenar los metadatos de *Datos Abiertos* y generar métricas de estos valores y buscar relaciones con la percepción de utilidad de los usuarios por medio de un mecanismo de evaluación.

CAPÍTULO 3

Desarrollo del Proyecto

3.1. Metodología

La metodología que se siguió en este trabajo de grado fué basada en los lineamientos de Desing Science Research (DSR) (vom Brocke et al., 2020), la cual propone solucionar problemas mediante el uso de artefactos innovadores que mejoran las condiciones y el entorno en el cual son instanciados. El resultado esperado de la metodología es tanto diseño de nuevos artefactos como nuevo conocimiento (vom Brocke et al., 2020).

De manera que, esta metodología se alinea con los objetivos propuestos en este trabajo de grado que plantea la creación de una herramienta de software para brindar indicadores para evaluar los datos disponibles de la plataforma *datos abiertos* del Gobierno Colombiano. Y siguiendo con sus lineamientos se describen las siguientes actividades:

1. Obj: Entender los metadatos disponibles de la plataforma *datos abiertos* y plantear indicadores para el proceso de evaluación.

    a) Identificar la estructura y la disposición de los metadatos.

    b) Analizar los metadatos y seleccionar cuáles son relevantes para determinar la utilidad de los datos.

2. Obj: Diseñar y construir un modelo dimensional para almacenar los metadatos relevantes para el análisis.

    a) Definir la topología del modelo que se usara para el Data Mart.

    b) Identificar las dimensiones del modelo.

    c) Identificar los hechos del modelo.

    d) Implementar el modelo en un Data Mart.

3. Obj: Implementar un proceso de extracción de los metadatos de la plataforma *datos abiertos* para almacenarlos en el modelo dimensional.

    a) Identificar la fuente para la extracción de los metadatos.

    b) Establecer el formato de la fuente y el formato de la salida del proceso.

    c) Diseñar e implementar el proceso de transformación.

4. Obj: Construir visualizaciones de los indicadores para evaluar la utilidad de los datos disponibles a partir de los metadatos.

   a) Seleccionar la herramienta de visualización.

   b) Configurar la herramienta de visualización y conectarla al Data Mart.

   c) Construir los tableros usando los gráficos correctos dependiendo de la información que se desea visualizar.

5. Obj: Evaluar la percepción de utilidad de la herramienta propuesta.

   a) Definir un instrumento de medición de la percepción de utilidad.

   b) Construir el instrumento.

   c) Ejecutar la medición.

   d) Procesar los resultados de la medición.

   e) Generar las conclusiones.

3.2. Entender los metadatos disponibles de la plataforma datos abiertos y plantear indicadores para el proceso de evaluación

3.2.1. Identificar la estructura y la disposición de los metadatos

La plataforma *datos abiertos* del Gobierno Nacional de Colombia se encuentra contenida dentro de otra plataforma mas grande conocida como *Socrata*, la cual expone una serie de servicios web para acceder a la información de la bodegas de datos que contiene, haciendo uso de estos servicios web, se obtiene la información de los metadatos de la plataforma *datos abiertos*.

Especificando la identificación *"njuy-iarc"* que corresponde a la plataforma *datos abiertos* dentro de *Socrata* y usando un token que se obtiene mediante el registro dentro de *Socrata* para desarrolladores se puede hacer uso de los metodos *HTTP* dispuestos para interactuar con la bodega de *datos abiertos* que es el objeto de estudio de este trabajo de grado.

Previamente se seleccionó un lenguaje de programación para hacer el respectivo servicio web y de esta manera consultar los metadatos. En este caso particular usaremos *Python* como lenguaje para acceder a los metadatos, debido a que es un lenguaje versatil y de facil uso, ademas *Socrata* proporciona una libreria para hacer la interacción mas eficiente.

Los elementos necesarios para este proceso se muetran a continuación en la Tabla 3.1:

Url:	“www.datos.gov.co"
Token:	“kD214IjFxxCjzOlETizfAUN3K"
Usuario:	“diegosegura448@gmail.com"
Password:	“870104Socrata–"
Identificación Bodega:	“njuy-iarc"

Tabla 3.1: Elementos para establecer la conexión con Socrata. Fuente: Elaboración propia.

Luego de implementar el metodo para la consulta del endpoint de *Socrata*, el segmento de código encargado de hacer la conexión se muesta en la figura 3.1.

```
import pandas as pd
from sodapy import Socrata

import json


client = Socrata("www.datos.gov.co",
                 "kD214IjFxxCjzOlETizfAUN3K",
                 username="diegosegura448@gmail.com",
                 password="870104Socrata--")
```

Figura 3.1: Fragmento de código que se conecta a *Socrata*. Fuente: Elaboración propia

Una vez generada la conexión se procedió a acceder a la bodega de *datos abiertos* usando la identificación que corresponde, como se muestra en la figura 3.2.

```
results = client.get("njuy-iarc", limit=40000)
```

Figura 3.2: Fragmento de código que usa el método get para obtener los metadatos. Fuente: Elaboración propia

3.2.2. Analizar los metadatos y seleccionar cuáles son relevantes para determinar la utilidad de los datos

Una vez que se obtuvieron los metadatos se inició la revisión para determinar cuáles de estos podrían aportar algun valor para determinar utlidad en los datos de la plataforma *datos abiertos*. Estos metadatos están compuestos por un total de 19 columnas, que se describen a continuación en la Tabla 3.2.

Metadato	**Descripción**
name:	Indica el nombre que se le proporciona al dataset.
description:	Breve descrición del contenido del dataset.
owner:	Institución o organismo al que pertenece el dataset.
publication_stage:	Estado en el que se encuentra la publicación del dataset.
type:	Tipo de dataset *"Dataset, federated_href, href, chart, map, filter, file, story, visualization, datalens, forms, calendar"*
category:	Indica a que categoria o rama del conocimiento pertenece el dato ver Tabla 3.3
creation_date:	Fecha de creación del dataset en la plataforma.
last_data_updated_date:	Fecha de última actualización del dataset.
approval_status:	Estado de aprovación del dataset, relacionado con campo "publication_stage".
visits:	Cantidad de visitas que tiene el dataset.
downloads:	Cantidad de descargas que tiene el dataset.
derived_view:	Indica si el dataset tiene información derivada, como otras vistas ó información adicional.
license:	Indica bajo que licencia se encuantra el dataset *"Creative Commons Attribution | NoDerivatives 4.0 International License, Share Alike 4.0 International, Noncommercial 3.0 Unported, Public Domain, SC, Creative Commons 1.0 Universal (Public Domain Dedication), Open Data Commons Public Domain Dedication and License"*.
informacindelaentidad_municipio:	Nombre del municipio o ente territorial que genera el dataset.
informacindelaentidad_nombredelaentidad:	Nombre de la entidad ó institución que genera el data set.
informacindelaentidad_departamento:	Nombre del departamento desde el que se genera el dataset.
row_count:	Cantidad de filas con las que cuenta el dataset.
contact_email:	Email de contacto de quien genera el dataset.
column_count:	Cantidad de columnas que tiene el dataset.

Tabla 3.2: Metadatos de la plataforma *datos abiertos* del Gobierno Nacional de Colombia. Fuente: Elaboración propia.

Valores metadato Categoria
Agricultura y Desarrollo Rural
Ambiente y Desarrollo Sostenible
Ciencia, Tecnología e Innovación
Comercio, Industria y Turismo
Cultura
Deporte y Recreación
Economía y Finanzas
Educación
Estadísticas Nacionales
Función pública
Gastos Gubernamentales
Hacienda y Crédito Público
Inclusión Social y Reconciliación
Justicia y Derecho
Mapas Nacionales
Minas y Energía
Organismos de Control
Presupuestos Gubernamentales
Salud y Protección Social
Seguridad y Defensa
Trabajo
Transporte
Vivienda, Ciudad y Territorio

Tabla 3.3: Valores del metadato Categoria, presentes en la plataforma de datos abiertos. Fuente: Elaboración propia.

Dado que la temática es responder si los datos son útiles basados en la información que se obtiene de los metadatos, inicialmente se seleccionan los metadatos que representan cantidades, y que pueden dar indicios de uso de manera intuitiva, como lo son los que se muestran a continuación en la Tabla 3.4.

Metadato
visits
downloads
row_count
column_count

Tabla 3.4: Metadatos candidatos para medir la utilidad. Fuente: Elaboración propia.

También pensando en como agrupar estos metadatos como indicadores que serán usados para visualización se seleccionaron los que se muestran en la Tabla 3.5.

Metadato
category
type
creation_date
last_data_updated_date
informacindelaentidad_municipio
informacindelaentidad_departamento

Tabla 3.5: Metadatos candidatos para medir la utilidad. Fuente: Elaboración propia.

A continuación se procedio a darle contexto a estos metadatos diseñando el modelo dimensional para el data mart que se propuso como parte de la solución, teniendo en cuenta la temática y el propósito que debe tener dicho data mart.

3.3. Diseñar y construir un modelo dimensional para almacenar los metadatos relevantes para el análisis

3.3.1. Definir la topología del modelo que se usara para el Data Mart

Basados en la temática del datamart la cual es identificar indicadores que puedan ayudar a determinar la percepción de utilidad, y basados en la cantidad de metadatos que provee la plataforma *datos abiertos*, se decidió utilizar la topología en copo de nieve, dado que una tabla de hechos puede no ser suficiente para poder almacenar los indicadores que nos aporten valor para la medición, en este caso se optó por generar 2 tablas de hechos para agrupar perfiles de los datasets representados en los metadatos, acompañadas de las dimensiones necesarias, orientados siempre a la temática propuesta para esta bodega.

Estos perfiles fueron pensados haciendo categorías con los metadatos que tienen listas de valores, de forma que se combinaron para armar las categorías, los metadatos seleccionados para estas categorías fueron los que se describen a continuación en la Tabla 3.6.

Metadato
Type
License
Column_count
Row_count
Derived_view

Tabla 3.6: Metadatos seleccionados para formar categorías y perfilar los datos: Elaboración propia.

De esta forma se espera que a largo plazo con estos perfiles se pueda extraer más información y agrupar tendencias, y así entender mejor el compartimiento de los datos que se ingresan a la plataforma, con esto en mente la idea de este modelo dimensional es tanto recolectar información de los datos y sus características, como acumular información basada en estos perfiles para descubrir nueva información futura que pueda ser brindada por el comportamiento de los datos.

3.3.2. Identificar las dimensiones del modelo

Para el diseño del data mart se construyeron 4 dimensiones y 2 tablas estáticas para normalizar la dimensión Ubicación que se compone del departamento y el municipio.

La primera dimensión tiempo, se definió a nivel diario, para registrar los cambios que se efectúan en la carga de los datos para el modelo, esta dimesión se definó como se muestra a continuació en la Figura 3.3.

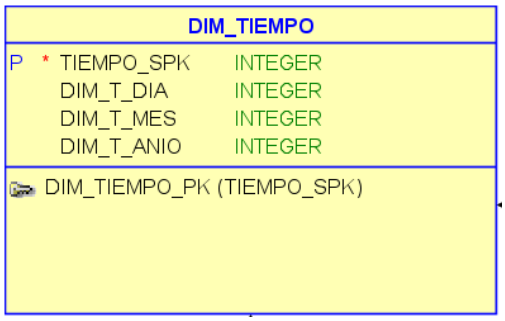


Figura 3.3: Dimensión Tiempo Modelo Dimensional. Fuente: Elaboración propia.

La dimensión dataset fué pensada para alojar las propiedades del dataset que si bien no brindan indicios de la utilidad, son necesarios para dar información particular de cada conjunto de datos, adicionalmente en el análisis previo de la información se contempló la idea de tener varias versiones de un dato, ya que existe un campo fecha que indica cuando fue la última vez que se actualizó el dato *"last_data_updated_date"*, de esta manera quedo definida la dimensión como se muestra a continuación en la Figura 3.4.

Posteriormente en ejecuciones hechas a través de la ejecución del experimento se notó que los datos a pesar de cambiar los valores de cantidad de columnas o cantidad de filas, como se esperaba, debían ser nuevas versiones, pero la plataforma registraba el mismo valor en las columnas de fecha de creación y de actualización, esto nos indicó que no se controla adecuadamente las fechas para tener un orden cronológico de cambios en los datos.

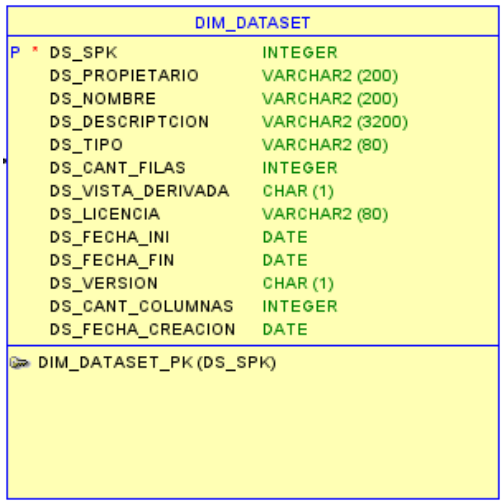


Figura 3.4: Dimensión Dataset Modelo Dimensional. Fuente: Elaboración propia.

Por otra parte la dimensión Ubicación se establece para indicar de dónde proviene el conjunto de datos en términos geográficos, ya que es relevante para dar contexto respecto de desde donde se están generando los datasets. Esta dimension se definió como se muestra en la Figura 3.5.

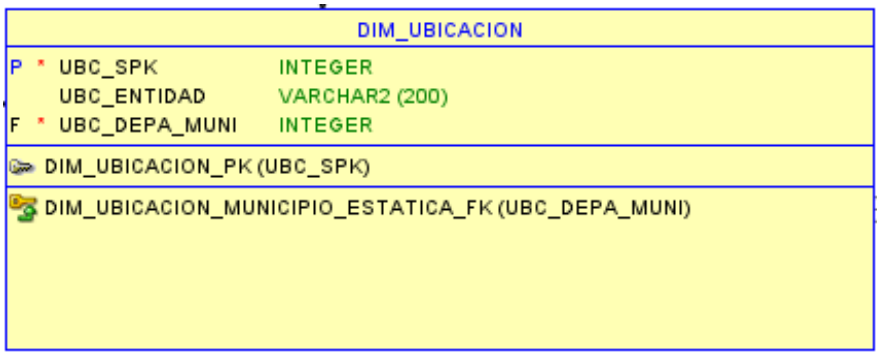


Figura 3.5: Dimensión Ubicación Modelo Dimensional. Fuente: Elaboración propia.

Adicionalmente se crearon dos tablas para estandarizar la información que proviene de la fuente, ya que está desorganizada y el usuario la genera, entonces contiene datos repetidos escritos de diferente manera en los nombres de los municipios y departamentos que se adicionan a los datasets las tablas fueron definidas como se muestra en la Figura 3.6 y Figura 3.7.

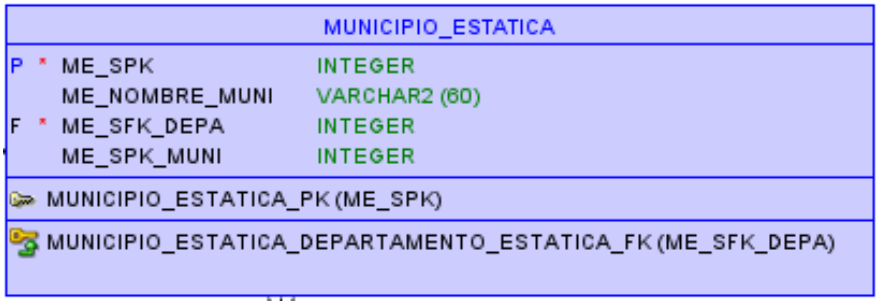


Figura 3.6: Tabla Municipio Modelo Dimensional. Fuente: Elaboración propia.

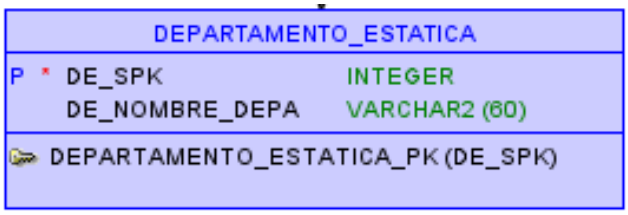


Figura 3.7: Tabla Departamento Modelo Dimensional. Fuente: Elaboración propia.

Finalmente se definió la dimensión perfil dataset con el fin de agrupar conjuntos de datasets y generar como su nombre lo indica, perfiles que ayudarán a encontrar comportamientos útiles para analizarlos y exponerlos en las visualizaciones. Esta dimensión se definió como se muestra acontinuación en la Figura 3.8.

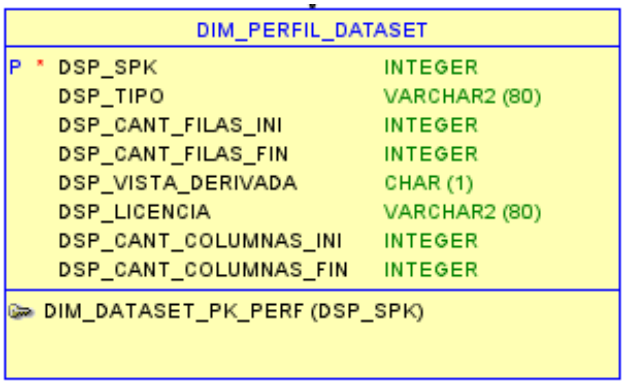


Figura 3.8: Dimensión Perfil Dataset Modelo Dimensional. Fuente: Elaboración propia.

3.3.3. Identificar los hechos del modelo

Para este modelo dimensional se definieron dos tablas de hechos, una de ellas para almacenar de forma particular las características de cada dataset como lo son *Descargas, Visitas y Categorías*, y la segunda donde se almacenaron las sumas de estas características. Con el fin de sacar los perfiles para el análisis de los comportamientos. Estas tablas de hechos fuerón definidas como se muestra en las Figuras 3.9 y 3.10 respectivamente.

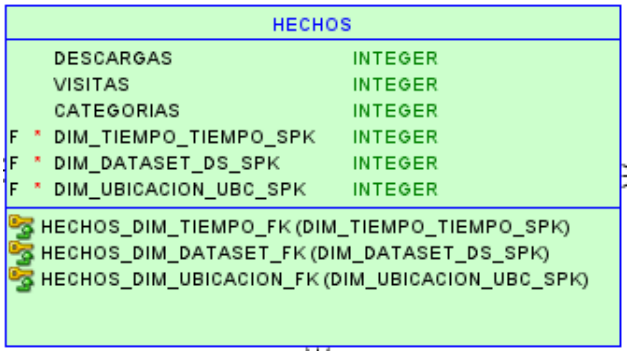


Figura 3.9: Tabla de Hechos Dataset Modelo Dimensional. Fuente: Elaboración propia.

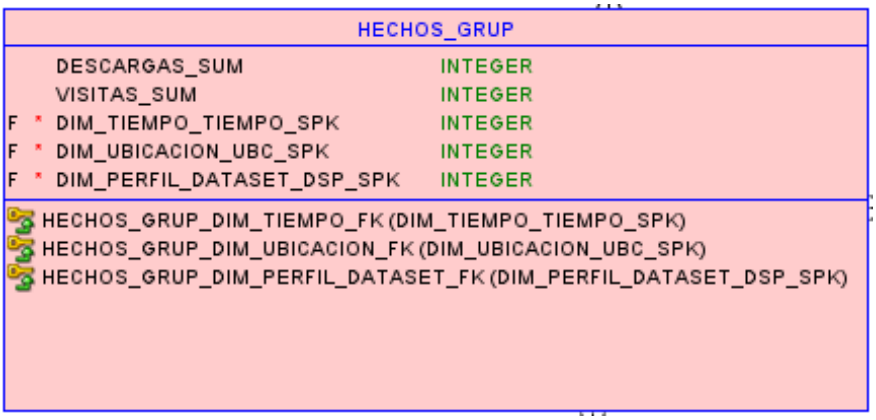


Figura 3.10: Tabla de Hechos Grupos Dataset Modelo Dimensional. Fuente: Elaboración propia.

Finalmente se utilizó una tabla pivote para poder realizar las transacciones que se ejecutan desde la ETL, esta tabla contiene la información que se genera después de que la transformación es ejecutada y su función es proveer la fuente de transacciones para el resto de tablas del Modelo Dimensional, se muestra a continuación en la Figura 3.11.

Figura 3.11: Tabla de Hechos Dataset Modelo Dimensional. Fuente: Elaboración propia.

Como resultado de lo mensionado, el modelo dimensional completo se muestra de la siguiente manera en la figura 3.12.

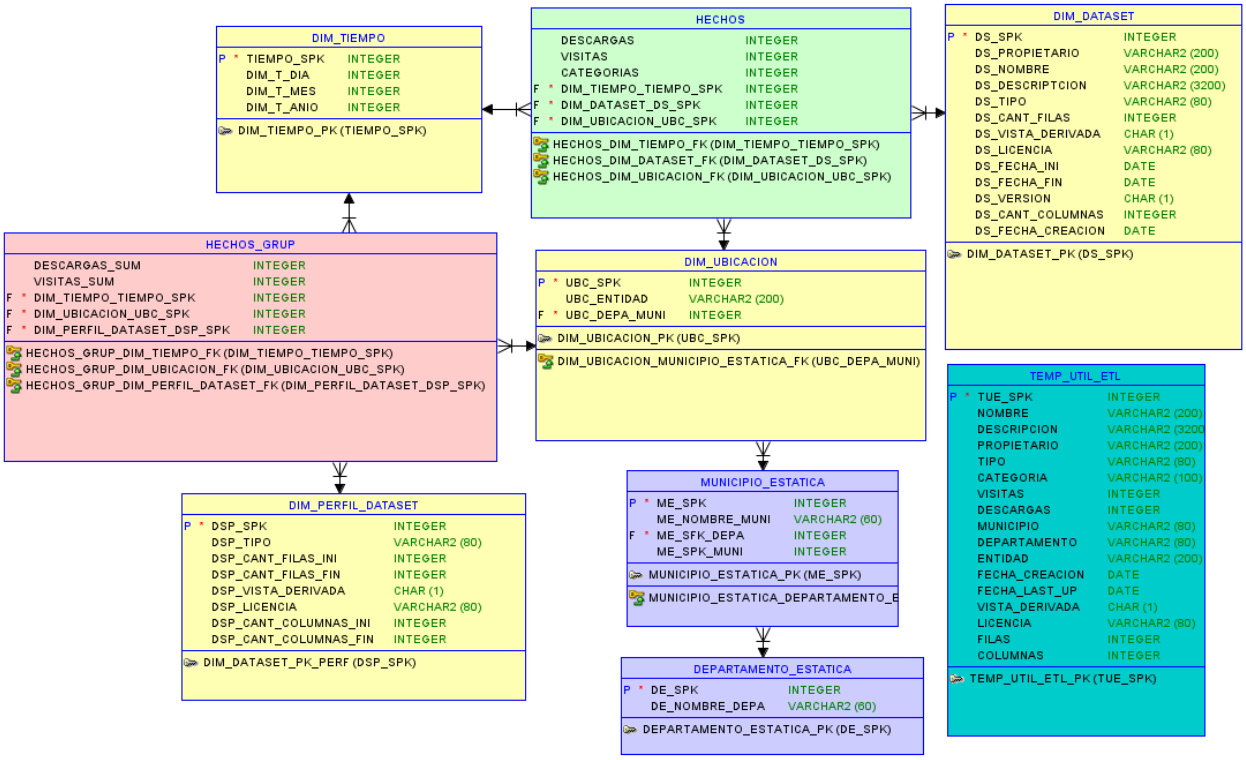


Figura 3.12: Modelo Dimensional. Fuente: Elaboración propia.

Partiendo de este modelo dimensional se inició con la elaboración de el proceso de *ETL (Extrac, Transform and Load)* que dispondrá los datos desde la fuente *Socrata* hasta el modelo dimensional propuesto, como se mostrará en la siguiente sección.

3.4. Implementar un proceso de extracción de los metadatos de la plataforma datos abiertos para almacenarlos en el modelo dimensional

3.4.1. Identificar la fuente para la extracción de los metadatos.

Para nuestro caso particular, el proceso de extracción tendrá que acceder a los servicios dispuestos por la plataforma *Socrata* para obtener los metadatos.

Socrata ofrece una gran cantidad de opciones para acceder a los datos de las bodegas que contiene, desde SDKs y APIs para diferentes lenguajes, hasta implementaciones particulares dependiendo de las necesidades de los usuarios.

En esta ocasión se usó una librería de *Python* llamada *sodapy* para acceder a la información que requerimos, y esta implementación está disponible en el repositorio de git que está compartido a través de la documentación de *Socrata* para desarrolladores (https://github.com/xmunoz/sodapy).

Estos datos desde el servicio están dispuestos en formato JSON inicialmente. Y serán procesados para hacer las correspondientes agrupaciones, además de la limpieza de los datos para garantizar que sean ingresados en el Data Mart.

3.4.2. Establecer el formato de la fuente y el formato de la salida del proceso.

Como se mencionó, inicialmente los metadatos serán transportados en formato JSON, pero para facilitar el manejo dentro del flujo de la ETL, esta información se transforma en formato csv para ser cargado en forma de tabla al inicio del proceso de transfomación, utilizando los conectores de la herramienta *Pentaho Data Integration (PDI)* bajo licencia community.

Luego de que el proceso finalice, los datos deben quedar almacenados en las tablas del modelo dimensional anteriormente expuesto para iniciar la fase siguiente que es el análisis e identificación de los perfiles para despúes presentar las representaciones gráficas a los usuarios para que estos hagan la evaluación correspondiente.

3.4.3. Diseñar e implementar el proceso de transformación.

Para este proceso de ETL se creó un Job con las transformaciones necesarias para insertar los datos correspondientes en el modelo dimensional descrito en la sección anterior, la estructura de este Job se muestra a continuación en la Figura 3.13.

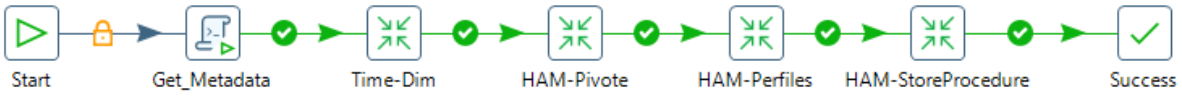


Figura 3.13: Job principal ETL. Fuente: Elaboración propia.

A continuación se describe cada una de las transformaciones mostradas en el Job.

Get_Metadata

Este paso es el encargado de conectarse a *Socrata* y extraer los datos mediante un script de *Python* y guardar la información en un archivo CSV para usarlo como fuente, ya que de esta forma se puede estandarizar los tipos de dato que en un principio vienen en formato Json desde el servicio de *Socrata*, pero contienen algunos errores de tipado con las cadenas y los números, el script se muestra a continuación en la Figura 3.14.

```
1         import pandas as pd
2         from sodapy import Socrata
3
4         client = Socrata("www.datos.gov.co",
5                          "kD214IjFxxCjzOlETizfAUN3K",
6                          username="diegosegura448@gmail.com",
7                          password="870104Socrata--")
8
9         results = client.get("njuy-iarc")
10        results_df = pd.DataFrame.from_records(results)
11        results_df.to_csv(".\Metadata.csv", index=False, encoding='utf-8')
```

Figura 3.14: Script para optener metadatos de *Socrata*. Fuente: Elaboración propia.

Time-Dim

En esta tranformación es la encargada de llenar la Dimensión del Tiempo, cada vez que el Job se ejecuta, toma el día, el mes y el año de la base de datos mediante el query, luego toma la secuecia de la tabla de la Base de Datos y la inserta. En la Figura 3.15 se ilustra.

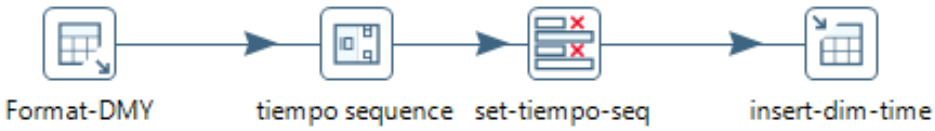


Figura 3.15: Tranformación Time-Dim. Fuente: Elaboración propia.

En el primer paso de la transformación se ejecuta el query que obtiene el día, el mes y el año de la Base de Datos, como se muestra en la Figura 3.16.

```
1         SELECT
2         EXTRACT(DAY FROM sysdate) AS dia,
3         EXTRACT(MONTH FROM sysdate) as mes,
4         EXTRACT(YEAR FROM sysdate) as anio
5         FROM DUAL
```

Figura 3.16: Query que optiene la fecha. Fuente: Elaboración propia.

Posteriomente se hace el llamado a la secuencia de la tabla correspondiente a la Dimensión Tiempo, para unificarlo con los datos de la fecha y finalmente hacer la inserción.

HAM-Pivote

Esta transformación es la encargada de procesar los datos del archivo CSV obtenido en el paso **Get Metadata**, filtrando los datos que necesitamos para el modelo y haciendo la respectiva modificación de los datos para su disposicion final en la tabla pivote, a continuación se ilustra esta transformación en la Figura 3.17.

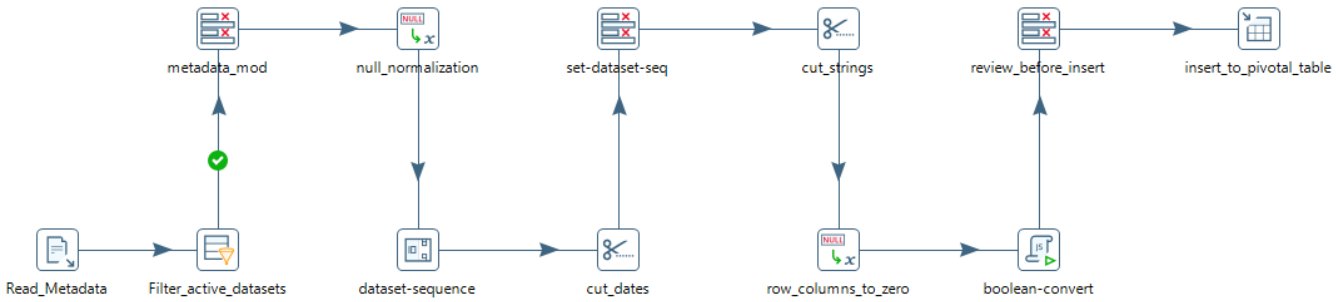


Figura 3.17: Tranformación HAM-Pivote. Fuente: Elaboración propia.

En el primer paso **Read_Metadata** se obtiene la información del archivo CSV, posteriormente en el paso **Filter_active_ datasets** se hace un filtrado de los datos que se encuentren activos con el campo *approval_status* con el valor *approved*, para no procesar registros que no se estén mostrando en *datos.gov*, seguidamente se asignan tipos de dato en el paso **metadata_mod**.

En el paso **null_normalization** se empezó a tomar decisiones respecto a los campos nulos que son de alguna manera importantes campos como *license, informacindelaentidad_departamento, informacindelaentidad_municipio y informacindelaentidad_nombredelaentidad*, que son necesarios para establecer la ubicación geográfica y el tipo de licencia bajo la cual se subieron los datos.

Con el fin de no perder datos por campos nulos, se decide reemplazar el dato nulo con el String “SC"para poder marcarlos y usar el resto de información útil que puedan tener, ya que si no se aplicaba ninguna estrategia se perdían al rededor de veinte mil datos.

Luego de solucionar el problema de los datos nulos, se prosiguió con la obtención de la secuencia de la tabla de hechos he incorporarla al flujo, otra corrección de los datos que se hizo, fue cortar las cadenas de fechas para darles formato tipo Date y cortar los campos *name* a 150 caracteres y *description* a 3.000 caracteres para estandarizar el tamaño de estas cadenas, ya que eran muy extensas.

Finalmente para los datos que tienen los campos *row_count y column_count* nulos se aplicó la estrategia de ponerlos en 0, ya que no todos los datos poseen estas columnas, como son los casos de los mapas, las ubicaciones o los gráficos. El último ajuste antes de la inserción fue normalizar el campo *derived_view_mod* cambiando los caracteres ’Y’ y ’N’ por valores 1 y 0.

HAM-Perfiles

Esta tranformación tiene como fin, insertar los datos necesarios en la dimensión **DIM_PERFIL_DATASET** que establece los rangos de los perfiles sobre los cuales se situaran los datasets para clasificarlos, y poder hacer un perfilamiento en la dimensión **Hechos Group** que funciona como una *Slowly Changing Dimention*, que es una de las formas en las que se reflejan los cambios de las fuentes de datos y estos cambios deben ser contenidos en este tipo de entidades para poder mostrarlos efectivamente.-(Santos and Belo, 2011). en la Figura 3.18 se ilustra la tranformación.

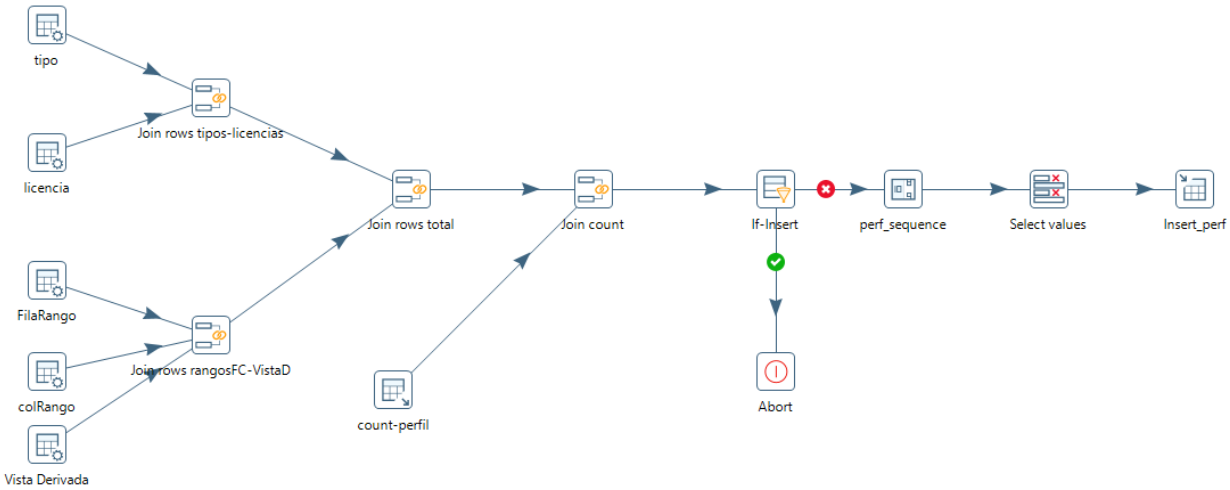


Figura 3.18: Tranformación HAM-Perfiles. Fuente: Elaboración propia.

Si miramos de derecha a izquierda la transformación, notamos que los primeros pasos **tipos, licencia, FilaRango, colRango y Vista Derivada** establecen los rangos de estos atributos para iniciar con el proceso de la convinación que dara lugar al perfil.

Luego usando los componentes **Join (Join rows rangosFC-VistaD, Join rows tipos-licencias y Join rows total)** para hacer la unión de estos valores y preparar el insert de la dimensión.

Como esta tarea de crear los perfiles se hace únicamente cuando se establece el Data Mart por primera vez, el componente **count-perfil** ejecuta un conteo a la Dimensión perfil para validar si esta vacía o no. Como se muestra en la siguiente Figura 3.19

```
1          SELECT count(*) as contar
2          FROM HAM.DIM_PERFIL_DATASET
```

Figura 3.19: Query que optiene el conteo de la dimensión perfil. Fuente: Elaboración propia.

Posteriormente en el paso **If-insert** se usa el resultado del conteo, si es 0, se procede con el siguiente paso, si es diferente de 0 entonces se procede a abortar el flujo y se continúa con la siguiente transformación.

En caso de que el contador esté en 0, se procede con el paso **perf_sequence** que genera el valor de la secuencias para cada valor que se va a insertar. EL paso **check_vals** es usado para validar visualmente una muestra en la ejecución y finalmente el paso **insert_perf** es el encargado de realizar el insert en la dimensión con los valores generados.

HAM-StoreProcedure

Esta transformación, solo contiene un paso **Call-StoreProcedure**. El llamado al procedimiento almacenado denominado **SP_HAM**, como se muestra en la Figura 3.23. El cual es el encargado de hacer el aprovisionamiento de los datos en todas y cada una de las dimensiones del modelo, cada vez que se ejecute la **ETL**. Realizando las respectivas actualizaciones e inserciones en el orden correcto, usando la tabla pivote para este fin.

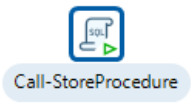


Figura 3.20: Tranformación HAM-StoreProcedure. Fuente: Elaboración propia.

Finalmente esta ETL fue programada para que se ejecute diariamente a las 5:00 AM (GTM-5), para que capture los cambios que se efectúen en la plataforma de datos abiertos. Para este fin se uso una tarea programada dentro del Sistema Operativo que ejecuta un script **schedule_HAMJob** con extensión **.bat** para windows y con extension **.sh** para ambientes linux, como se muestra en la Figura 3.21.

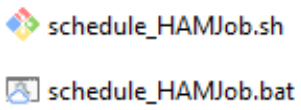


Figura 3.21: Scripts para ejcución automatica. Fuente: Elaboración propia.

Internamente las instrucciones de los scripts contienen un llamado a la aplicación **Kitchen** que hace parte de *Pentaho Data Integration*, estas instrucciones se muestran acontinuación en la Figura 3.22.

Al finalizar la implementación de la ETL y en las ejecuciones posteriores se notó que existían elemen-

```
1        cd ./pdi-ce-9.4.0.0-343/data-integration
2        kitchen.bat /file:"./samples/jobs/HAM-Jobs/Job_HAM.kjb"
```

Figura 3.22: Contenido de script **.bat** para llamado automático de la ETL. Fuente: Elaboración propia.

tos (datos) que eran idénticos, pero algunas características como el número de filas o número de columnas cambiaba, esto indicaba que era una versión nueva de dicho dato, sin embargo las fechas capturadas por la plataforma *Datos Abiertos* eran exactamente iguales a las fechas de creación del elemento. Lo que nos indicó que la plataforma no controla este tipo de escenarios.

Así mismo se evidenció que para muchos de los datos almacenados y provenientes de la plataforma *Datos Abiertos*, no hay una coherencia directa entre el nombre del dato y la descripción que los usuarios aportan al momento de compartir la información, y la plataforma tampoco posee control sobre este tipo de comportamiento, un ejemplo de esto se muestra en la siguiente imagen3.24

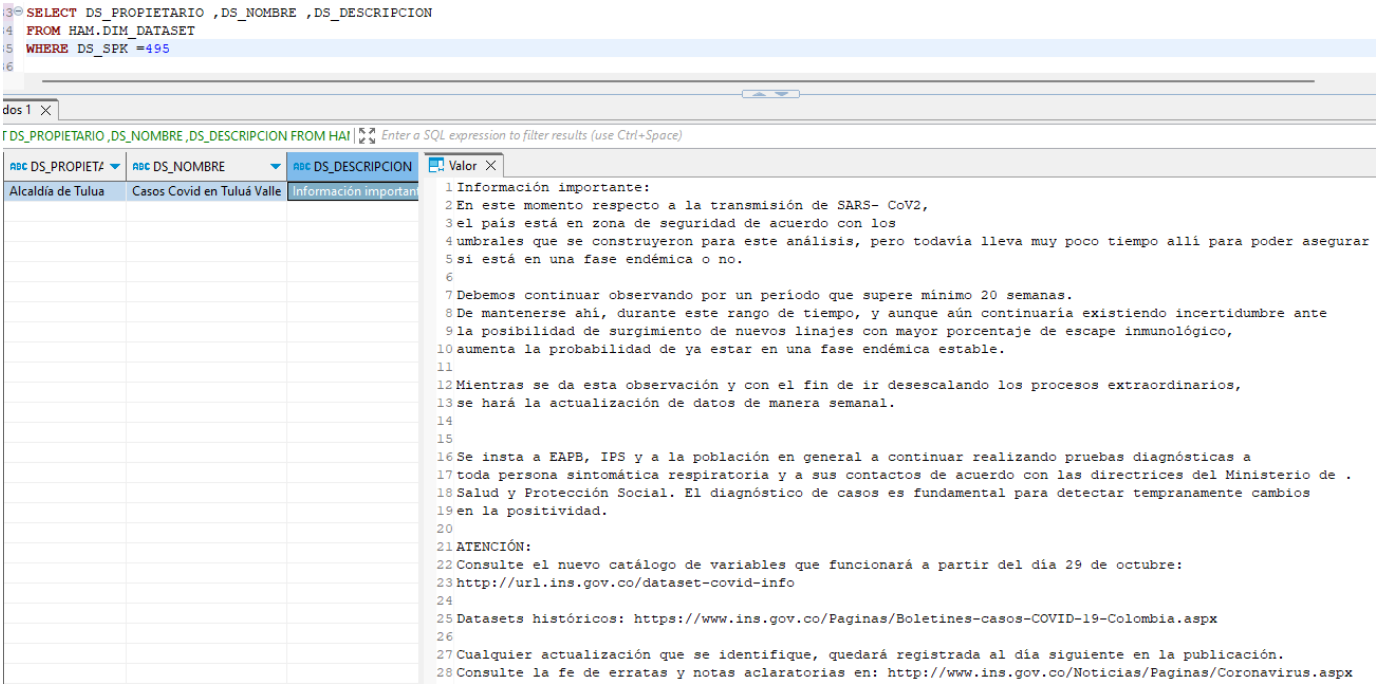


Figura 3.23: Ejemplo diferencia entre nombre del dato y la descripción. Fuente: Elaboración propia.

Una vez automatizada la ejecución de la ETL, se procedió a dar inicio con la visualizació de los datos, como se muestra en el siguiente sección.

3.5. Construir visualizaciones de los indicadores para evaluar la utilidad de los datos disponibles a partir de los metadatos.

3.5.1. Seleccionar la herramienta de visualización.

Para esta parte del trabajo se seleccionó la herramienta de visualización *Power BI*, ya que es una de las más usadas y se acopló a la necesidad del presente documento.

Esta herramienta cuanta con un amplio set de elementos útiles para trabajar con volúmenes de datos importantes, además de diferentes tipos de gráficos y herramientas colaborativas. Esto es un plus a la hora de compararla con otras herramientas como *Tableau*. Aunque *Power BI* tiene licenciamiento, para el uso que se le dió en el desarrollo de este trabajó solo se usó su capa gratuita.

3.5.2. Configurar la herramienta de visualización y conectarla al Data Mart.

Esta actividad se hizo usando el cliente de Oracle y el manual de conexión que ofrece *Power BI* para conectar instancias Oracle.

Una vez conectado se cargó la información del modelo con los datos dentro de la aplicación para poder iniciar con la etapa siguiente que es la de generar los tableros para que los usuarios pudieran visualizar la información que se obtuvo de los pasos anteriores de la ejecución del proyecto.

3.5.3. Construir los tableros usando los gráficos correctos dependiendo de la información que se desea visualizar.

En la construcción de los tableros, inicialmente se buscó hacer un análisis de los posibles indicadores relevantes que pudieran aportarle a los usuarios un criterio desde la mirada de los metadatos y así evaluar que tan útil para ellos es la plataforma de datos abiertos de Colombia.

De los metadatos disponibles, el metadato Descargas y el metadato Visitas son los que directamente nos pueden dar un indicio de utilidad más directo, por esto se evaluaron en conjunto con otros metadatos que pueden dar al Usuario herramientas para evaluar esta utilidad.

Con esto en mente los indicadores que se desarrollaron fueron:

Descargas vs Visitas, discriminados por categorias, como se muestra en la Figura 3.24.

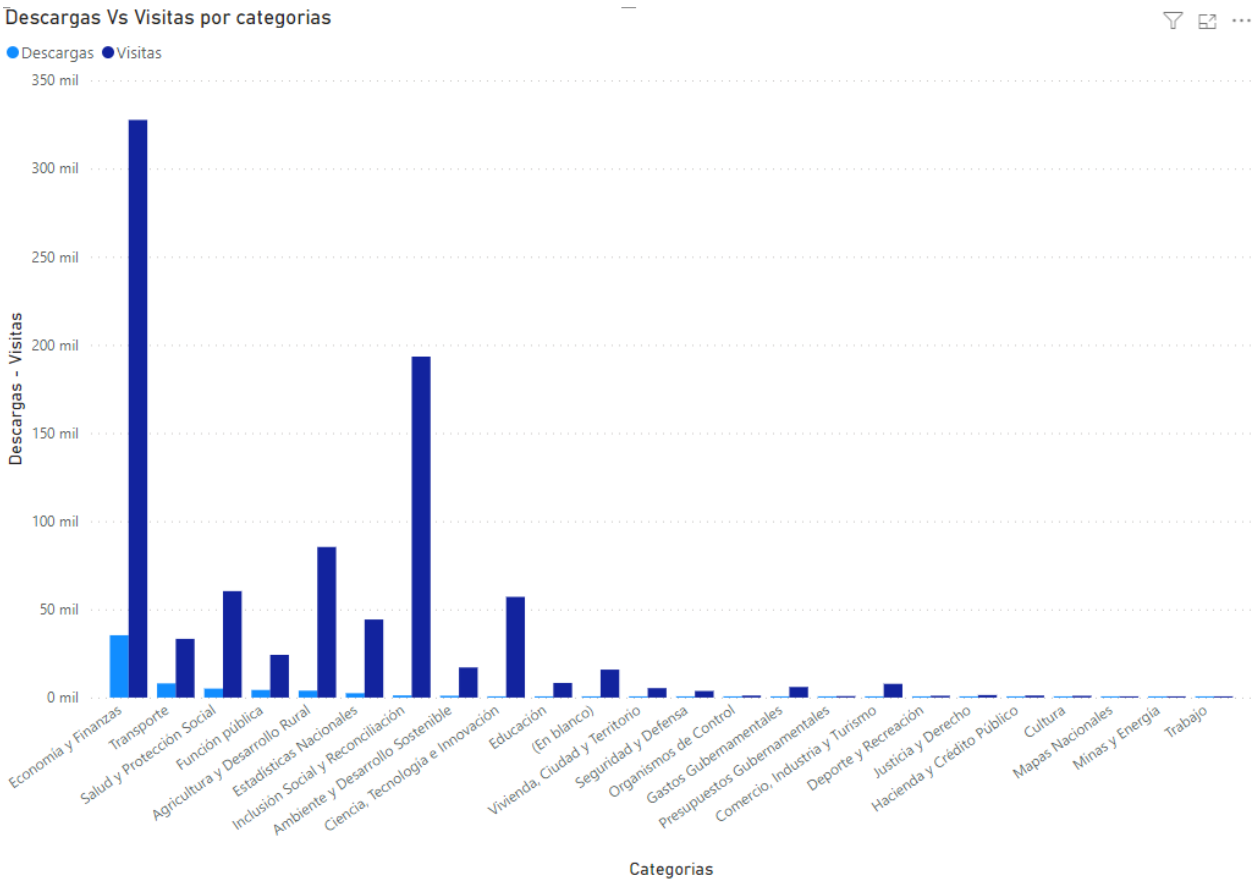


Figura 3.24: Indicator Descargas Vs Visitas por Categoria. Fuente: Elaboración propia.

Este indicador ilustra la relación que existe entre visitas y descargas por cada categoría disponible en la plataforma de datos abiertos, donde muestra una clara tendencia de que a pesar de ser muy visitada la categoría Economía y Finanzas con más de 300.000 visitas, los datos no son muy descargados con menos de 50.000 descargas.

Descargas por Departamento, como se muestra en la Figura 3.25.

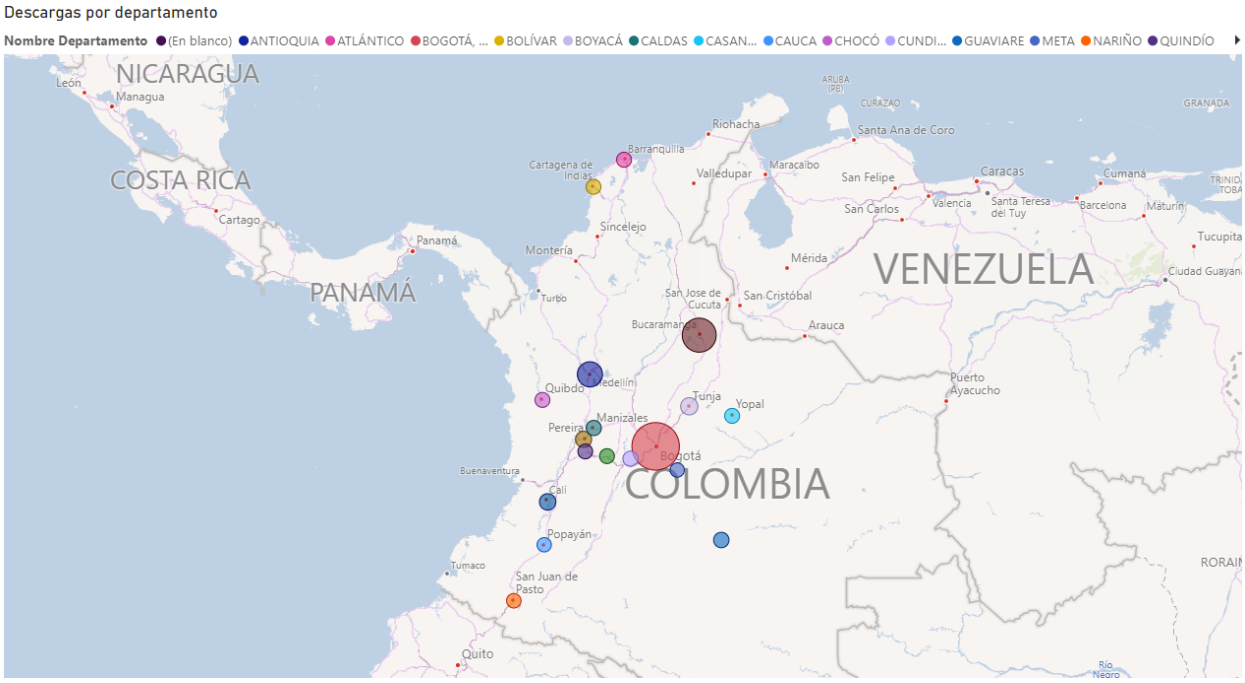


Figura 3.25: Indicator Descargas por Departamento. Fuente: Elaboración propia.

En el mapa se identifica que la zona con más descargas es Bogotá, con una diferencia considerable que observaremos más adelante usando un histograma.

Visitas por Departamento, como se muestra en la Figura 3.26.

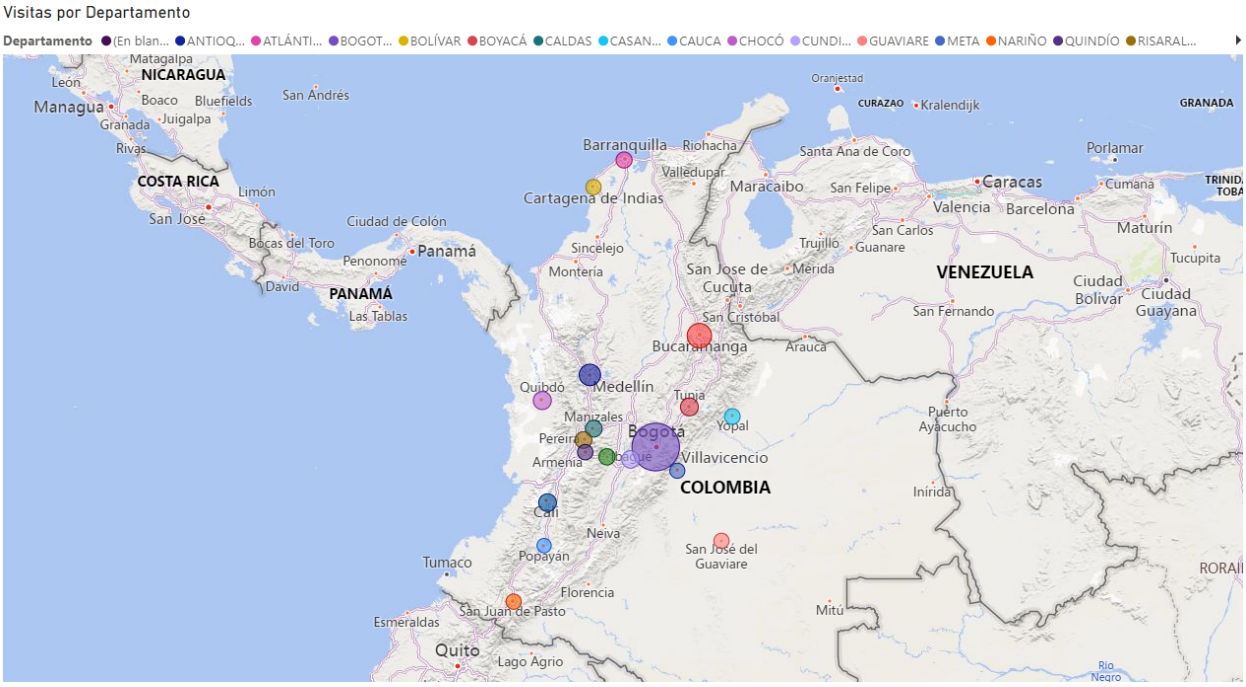


Figura 3.26: Indicator Visitas por Departamento. Fuente: Elaboración propia.

Se observa en el mapa que la Bogotá continúa teniendo ventaja frente a los demás departamentos en cuanto a las visitas, así como en descargas, donde también cuenta con un mayor número en comparación con los demás departamentos. A continuación se analizarán las descargas y visitas por departamentos para ver sus cantidades.

Descargas Vs Visitas por Departamento, como se muestra en la Figura 3.27.

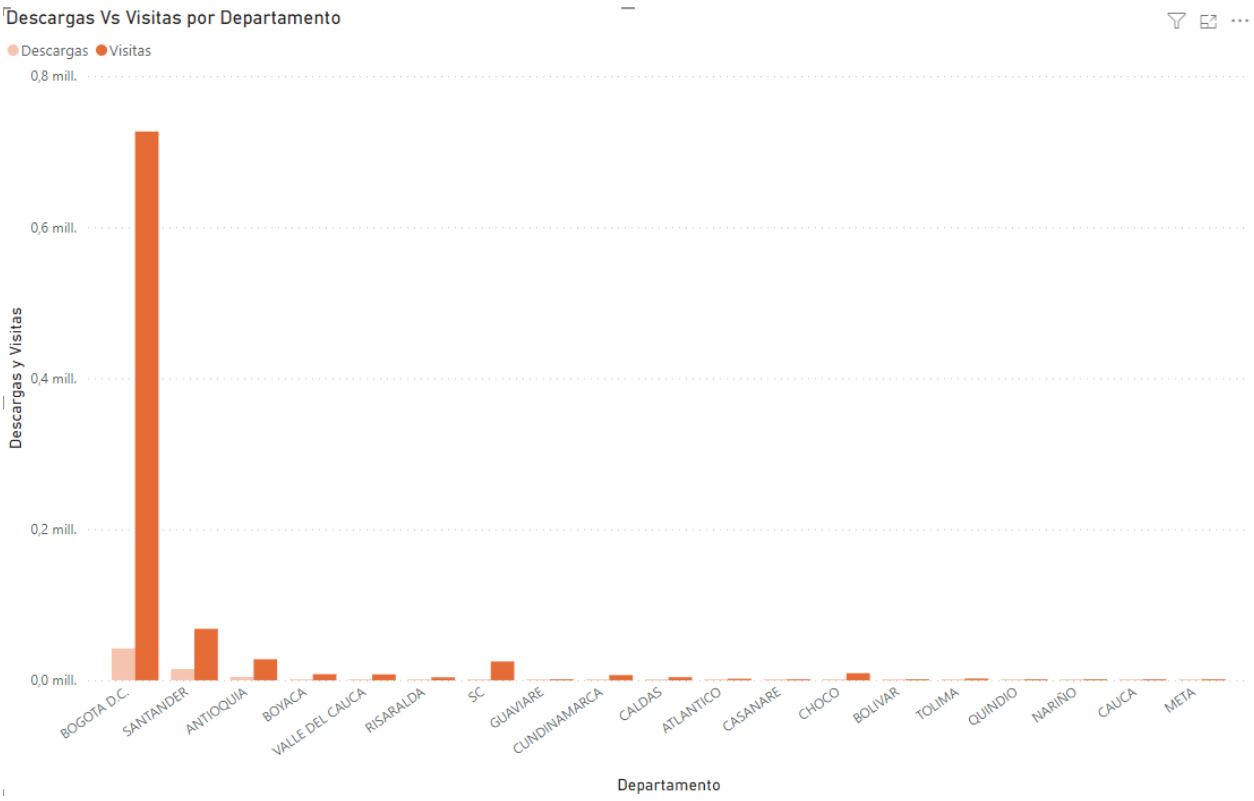


Figura 3.27: Indicator Descargas Vs Visitas por Departamento. Fuente: Elaboración propia.

En esta gráfica se ilustra las descargas y visitas por ubicación departamental, donde podemos observar que BOGOTÁ D.C cuenta con la mayor cantidad de visitas, con casi 800.000, sin embargo, menos de 50.000 descargas. En esta gráfica también se observa la ubiciación **SC** la cual se añadió para cubrir los registros que no cuentan con una ubicación desde la fuente de datos, son casi 30.000 registros que no se pueden identificar por departamento, aunque es una porción pequeña hay que tenerla encuenta, ya que la plataforma no hace ningún tipo de validación respecto a estos metadatos.

Descargas vs Visitas por Licenciamiento, como se ilustra en la Figura 4.1.

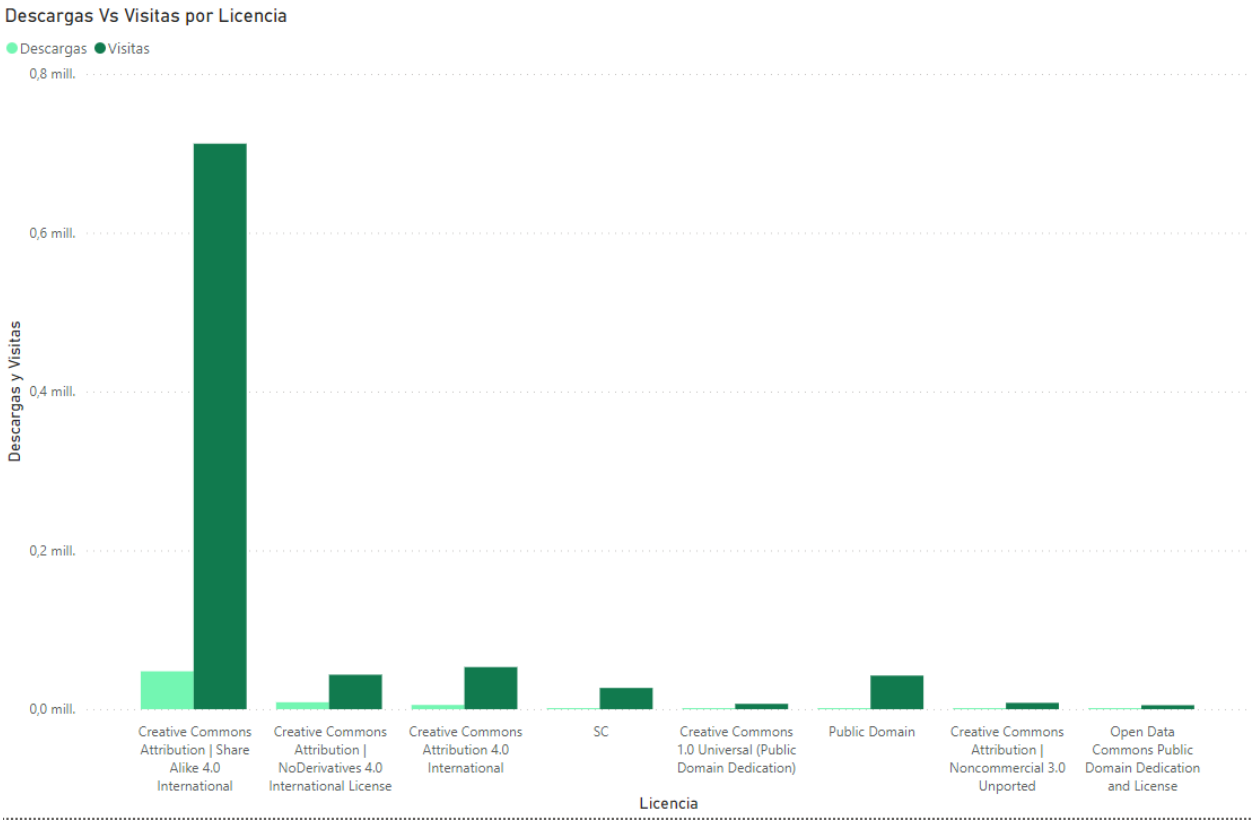


Figura 3.28: Indicator Descargas Vs Visitas por Licencia. Fuente: Elaboración propia.

Este indicador muestra la relación entre descargas y visitas pero por tipo de licenciamiento de los conjuntos de datos. Es un dato interesante, teniendo en cuenta que hablamos de una plataforma de datos abiertos, la cual debe promover el uso libre de los datos con el propósito de generar nuevo conocimiento, ya que es el eje principal de este movimiento. Se muestra que el tipo de licencia mas visitado es *Creative Commons Attribution | Share Alike 4.0 International* con casi 800.000 visitas. Sin embargo y siguiendo la misma tendencia pocas descargas con menos de 50.000.

Descargas vs Visitas por tipo de conjunto de dato, como se muestra en la Figura 4.2.

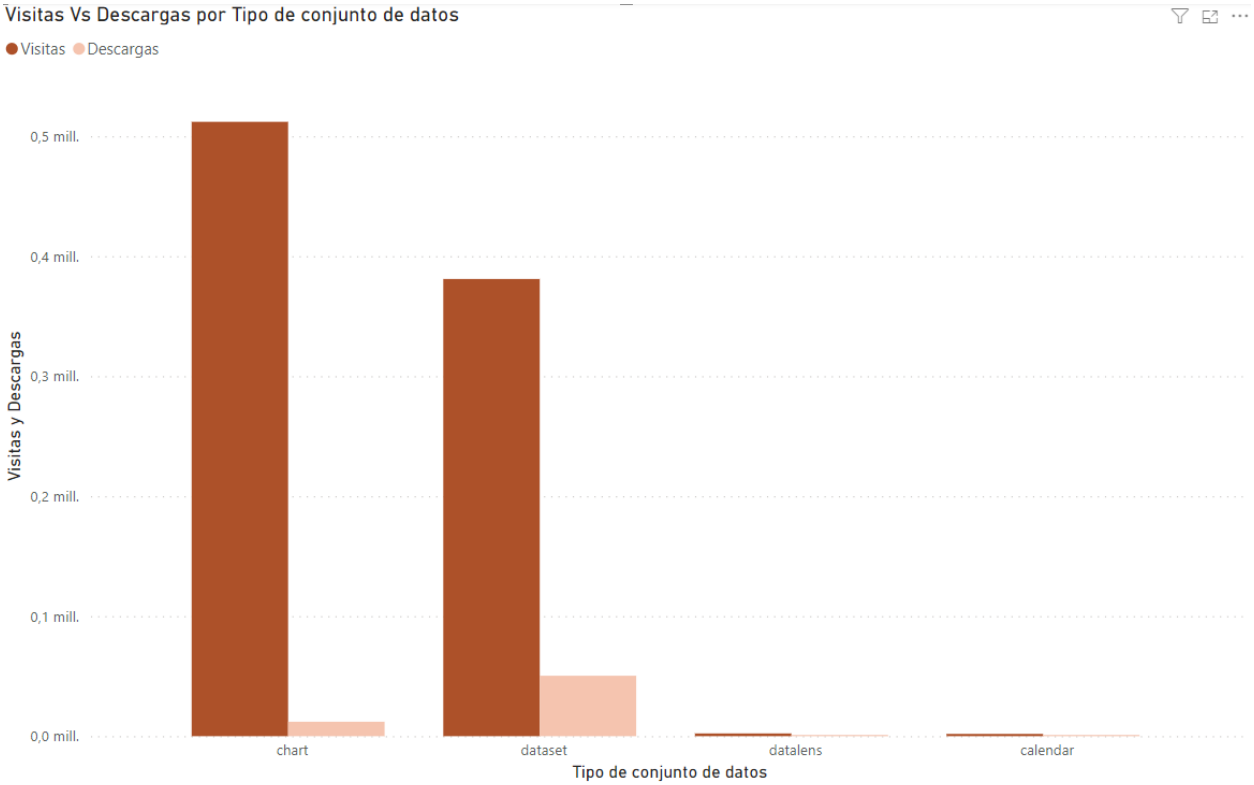


Figura 3.29: Indicator Descargas Vs Visitas por Tipo de conjunto de dato. Fuente: Elaboración propia.

Para este indicador, se pudo observar que hay 2 tipos de conjunto de datos que son los más visitados *chart* y *dataset*, el primero con un poco más de 500.000 visitas y el segundo con más de 400.000 visitas. Si embargo en cantidad de descargas es superior *dataset* con un número por encima de 50.000, seguido de *chart* apenas superando las 12.000.

Descargas vs Visitas por entidad que aporta los datos, como se muestra en la Figura 4.3.

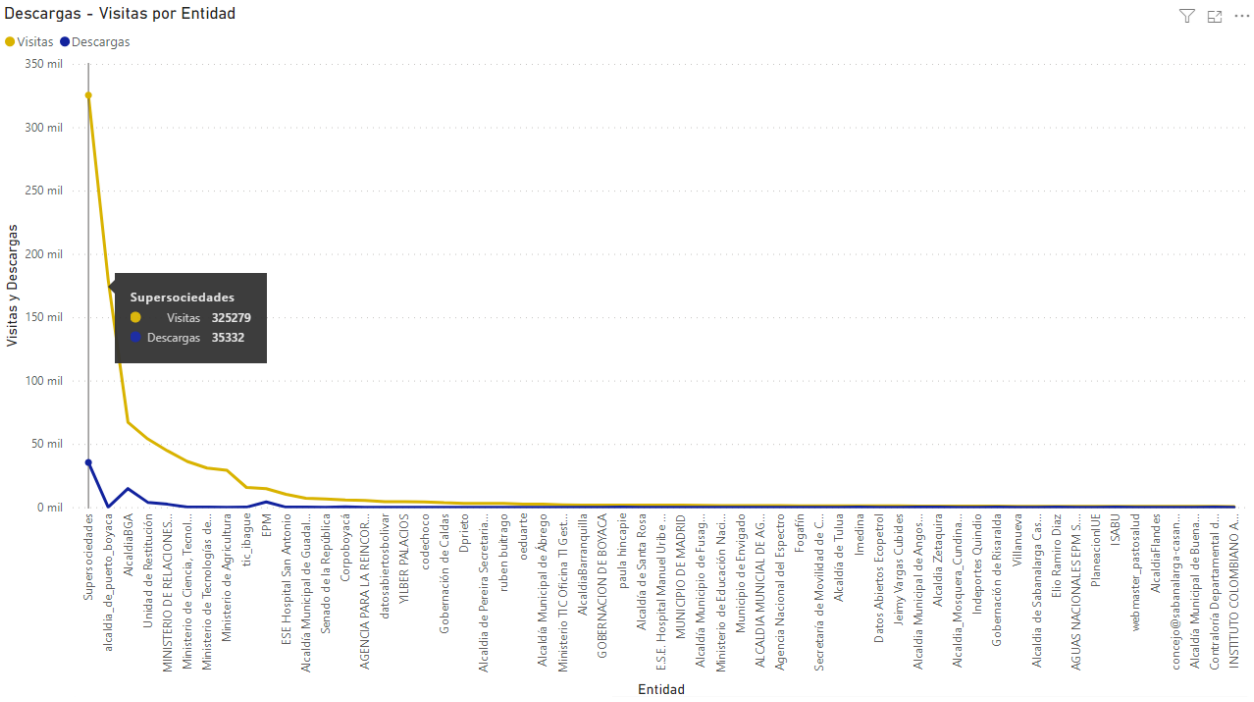


Figura 3.30: Indicator Descargas Vs Visitas por entidad que aporta los datos. Fuente: Elaboración propia.

En este indicador se muestra que la entidad que posee más descargas y visitas es supersociedades con 325.279 visitas y 35.332 descargas, muy por encima de la segunda entidad, alcaldía de Puerto Boyacá que cuenta con 179.447 visitas y apenas 8 descargas para sus datos.

Teniendo estos indicadores como insumo para que el usuario pueda indicar el nivel de utilidad se continúa con el siguiente punto, donde se formula el instrumento de medición que se usó para este fin.

3.6. Evaluar la percepción de utilidad de la herramienta propuesta

3.6.1. Definir un instrumento de medición de la percepción de utilidad

La percepción de utilidad es una medida que tiene fundamento teórico en variedad de estudios academicos, no solo relacionados a la tecnología sino a otras áreas del conocimiento. En este caso particular se hace uso de un cuestionario, para llevar a cabo la medición. Este cuestionario consta de preguntas que seran evaluadas para determinar la percepción de utilidad.

Esta serie de preguntas estan hechas basadas en teorias que buscan formular un cuestionario que sea fiable y valido (Heise and Bohrnstedt, 1970) para evaluar un aspecto del conocimiento, en este caso una herramienta de software, a continuación se muestran las preguntas para hacer la medición.

NOTA: *Se referenciará dentro del cuestionario la Herramienta para el análisis de los metadatos del portal Open Data Colombia con la abreviación "HAM"*

- ¿HAM Open Data Colombia optimiza la calidad del trabajo que realiza?
- ¿HAM Open Data Colombia incrementa su productividad?
- ¿HAM Open Data Colombia hace más fácil su trabajo?
- ¿En general HAM Open Data Colombia es útil en su trabajo?
- ¿Los indicadores mostrados por HAM Open Data Colombia son útiles para usted?
- ¿Los metadatos mostrados por HAM Open Data Colombia son suficientes para escoger un buen conjunto de datos?
- ¿Recomendaría los indicadores propuestos por HAM Open Data Colombia?
- ¿Con los indicadores mostrados, cree usted que HAM Open Data Colombia es útil?
- ¿Considera usted que se necesitan más métricas para evaluar la utilidad de HAM Open Data Colombia?

Se usó este conjunto de preguntas para la evaluación del concepto de percepción de utilidad de HAM Open Data Colombia.

3.6.2. Ejecutar la medición usando el instrumento

Para la evaluación de las preguntas se usó una escala tipo Likert (MoralesVallejoPedro, 2003), esta escala fué propuesta por el psicólogo *Rensis Likert*, ayuda a efectuar el análisis y la evaluación del cuestionario usando una escala de valores que el cuestionado, puede asimilar de forma uniforme, ya que las distancias entre cada elemento de la respuesta son iguales. Valores que se ilustran en la tabla 3.7:

Escala Likert
Totalmente en desacuerdo
En desacuerdo
Ni de acuerdo ni en desacuerdo
De acuerdo
Totalmente de acuerdo

Tabla 3.7: Valores de respuesta para preguntas del instrumento de medición.

Los cuestionados fueron encontrados haciendo uso de redes sociales y en el ámbito empresarial, buscando que conocieran de antemano la plataforma datos.gov.co. Dado que es de vital importancia para este trabajo que las personas a quienes se les pidió usar el instrumento tengan este conocimiento previo y así poder evaluar con mayor precisión.

El perfil de los encuestados es de profesionales con experiencia en datos, data scientist, ingenieros de datos, que trabajen con datos o en áreas relacionadas al análisis, y tratamiento de datos.

Además que cuenten con experiencia en el sector empresarial o docente, y que hayan interactuado con la plataforma. Usando estos criterios se encontró una muestra de 15 expertos que evaluaron la plataforma a través del instrumento de medición planteado anteriormente.

En el siguiente capítulo se mostrará como se hizo el análisis de los resultados que arrojo el instrumento de medición formulado al grupo de cuestionados.

CAPÍTULO 4

Evaluación

4.1. Análisis

4.1.1. Anotaciones iniciales

Una vez aplicado el instrumento de medición propuesto, se inició un proceso de evaluación de los resultados arrojados por el instrumento, a partir de este se desarrolló lo siguiente.

Para realizar una evaluación númerica, se asignó un número a cada pregunta para poder referirse a ella y medirla como se muestra en tabla 4.1:

Valor asignado	Pregunta
1	¿HAM Open Data Colombia optimiza la calidad del trabajo que realiza?
2	¿HAM Open Data Colombia incrementa su productividad?
3	¿HAM Open Data Colombia hace más fácil su trabajo?
4	¿En general HAM Open Data Colombia es útil en su trabajo?
5	¿Los indicadores mostrados por HAM Open Data Colombia son útiles para usted?
6	¿Los metadatos mostrados por HAM Open Data Colombia son suficientes para escoger un buen conjunto de datos?
7	¿Recomendaría los indicadores propuestos por HAM Open Data Colombia?
8	¿Con los indicadores mostrados, cree usted que HAM Open Data Colombia es útil?
9	¿Considera usted que se necesitan más métricas para evaluar la utilidad de HAM Open Data Colombia?

Tabla 4.1: Asignación de valores numéricos a las preguntas del instrumento para evaluarlas. Fuente: Elaboración propia.

Se hizo lo mismo con las respuestas, dándoles un peso a cada valor de respuesta del 1 al 5 como se ilustra en la tabla 4.2, dado que nos interesa saber que tan de acuerdo está el usuario con la herramienta

propuesta para poder determinar si es útil o no este análisis de la plataforma **datos.gov.co** a través de sus metadatos.

Valor asignado	Respuesta
1	Totalmente en desacuerdo
2	En desacuerdo
3	Ni de acuerdo ni en desacuerdo
4	De acuerdo
5	Totalmente de acuerdo

Tabla 4.2: Asignación de valores numéricos a las respuestas del instrumento para evaluarlas. Fuente: Elaboración propia.

4.1.2. Análisis de frecuencia

A continuación se muestran las frecuencias de las respuestas por pregunta en la tabla 4.3, usando los valores asignados a las preguntas en la columna pregunta.

Pregunta	Totalmente en desacuerdo	En desacuerdo	Ni de acuerdo ni en desacuerdo	De acuerdo	Totalmente de acuerdo
1	0	2	3	8	2
2	0	2	5	6	2
3	0	1	8	4	2
4	1	2	6	5	1
5	0	1	6	4	4
6	1	0	6	6	2
7	0	1	7	4	3
8	1	1	6	5	2
9	1	0	8	5	1

Tabla 4.3: Frecuencias de respuestas por pregunta. Fuente: Elaboración propia.

La siguiente Figura 4.1 muestra los resultados de la tabla anterior gráficamente, donde se puede observar que en términos generales las respuestas **De acuerdo** y **Ni en desacuerdo ni de acuerdo** obtuvieron la mayor cantidad de respuestas.

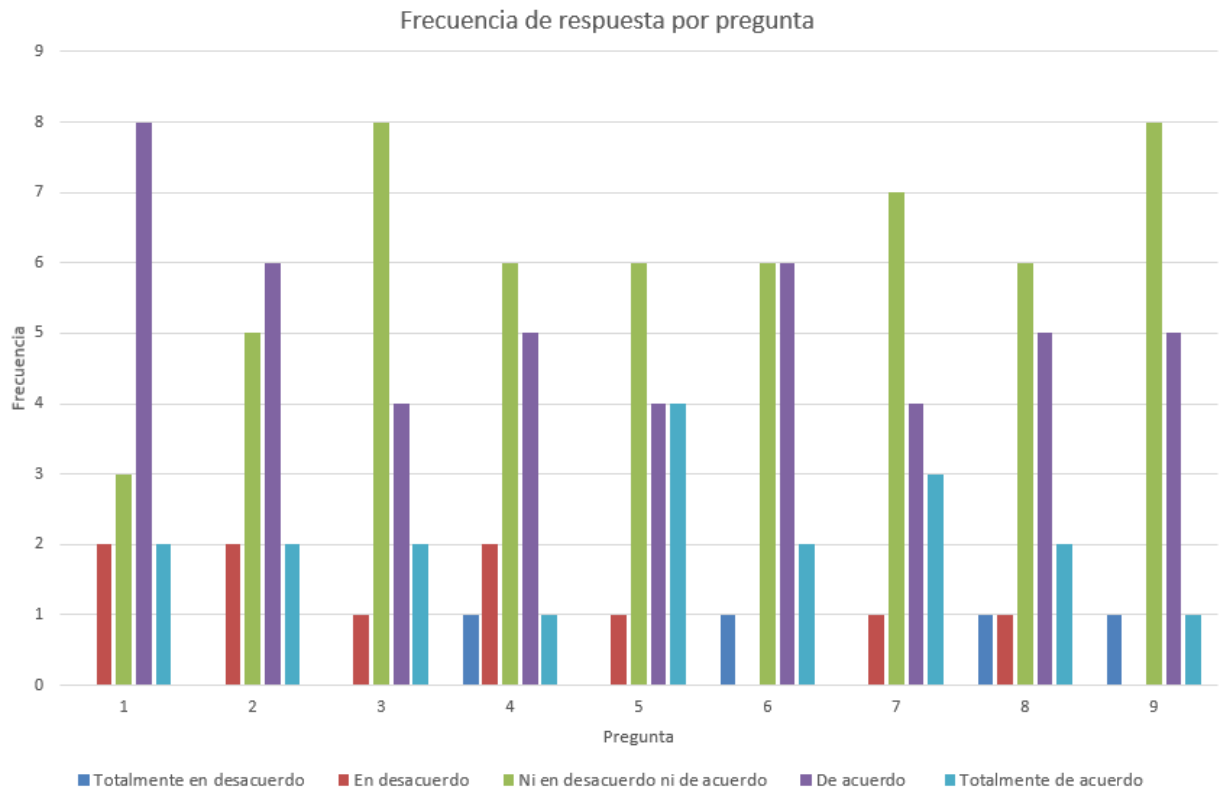


Figura 4.1: Gráfica que representa la frecuencia por respuesta.. Fuente: Elaboración propia.

Como se mencionó, las respuestas **De acuerdo** y **Ni en desacuerdo ni de acuerdo** muestran una clara diferencia con el resto, lo que inicialmente nos puede dar un buen indicio del comportamiento general de la medición, se continuó con un análisis más detallado por cada pregunta, para validar sus porcentajes en la siguiente sección.

4.1.3. Análisis de frecuencia por pregunta

También se hizó un análisis de la frecuencia por pregunta, en terminos de porcentaje, dando como resultado los siguientes porcentajes por pregunta:

Para la pregunta número 1, el 53.3 % de las respuestas manifiestan estar de acuerdo en que HAM optimiza la calidad de su trabajo, como se muestra en la Figura 4.2:

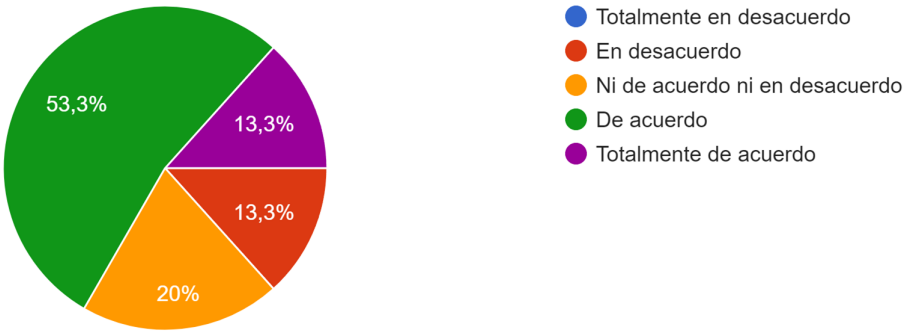


Figura 4.2: Resultado de la pregunta, ¿HAM Open Data Colombia optimiza la calidad del trabajo que realiza?. Fuente: Elaboración propia.

Para la pregunta número 2 se observa que el 40 % esta de acuerdo en que HAM incrementa su productividad, como se muestra en la Figura 4.3.

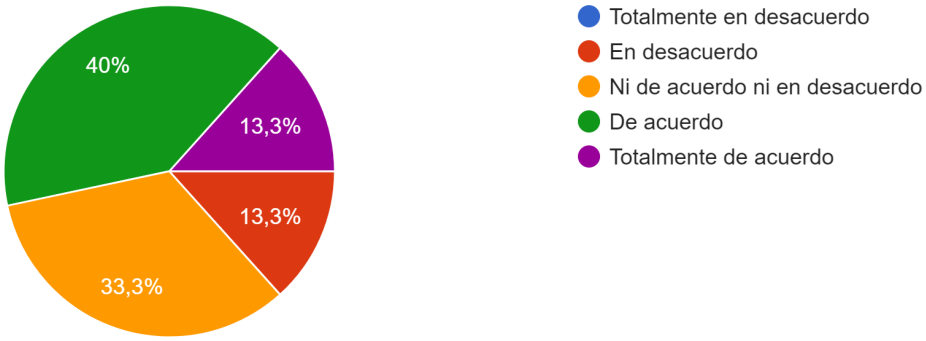


Figura 4.3: Resultado de la pregunta, ¿HAM Open Data Colombia incrementa su productividad?. Fuente: Elaboración propia.

Para la pregunta número 3 se muestra en la Figura 4.4 que el 53.3 % de las respuestas no esta ni en desacuerdo ni deacuerdo en que HAM hace mas facil su trabajo.

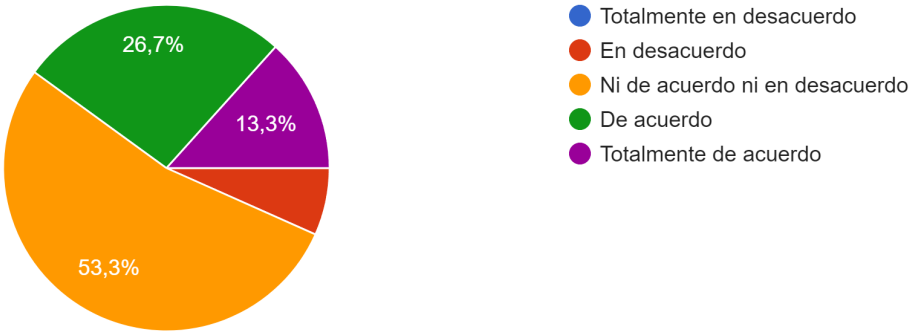


Figura 4.4: Resultado de la pregunta, ¿HAM Open Data Colombia hace más fácil su trabajo?. Fuente: Elaboración propia.

Se observa en la pregunta número 4 que el 40 % no esta ni en desacuerdo ni de acuerdo con que HAM sea útil para su trabajo, sin embargo seguido por un 33.3 % que esta de acuerdo con esa pregunta, como se ilustra en la Figura 4.5.

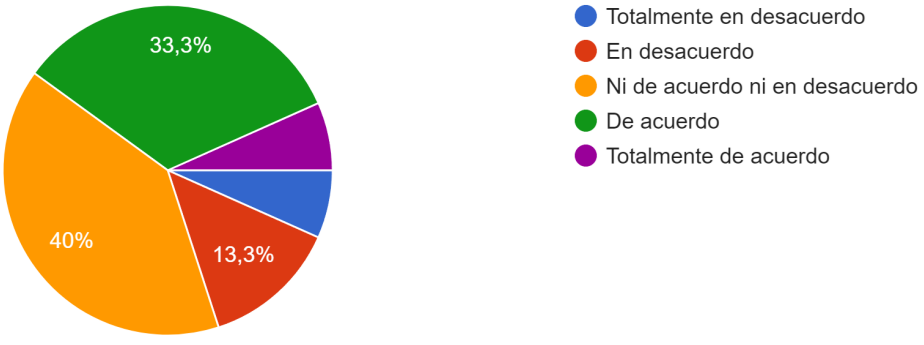


Figura 4.5: Resultado de la pregunta, ¿En general HAM Open Data Colombia es útil en su trabajo?. Fuente: Elaboración propia.

Para la pregunta número 5 se encuentra que un 40 % no esta ni en desacuerdo ni de acuerdo en que los indicadores de HAM sean utiles para ellos, seguido de un 26.7 % que esta de acuerdo y totalmente de acuerdo, como se muestra en la Figura 4.6.

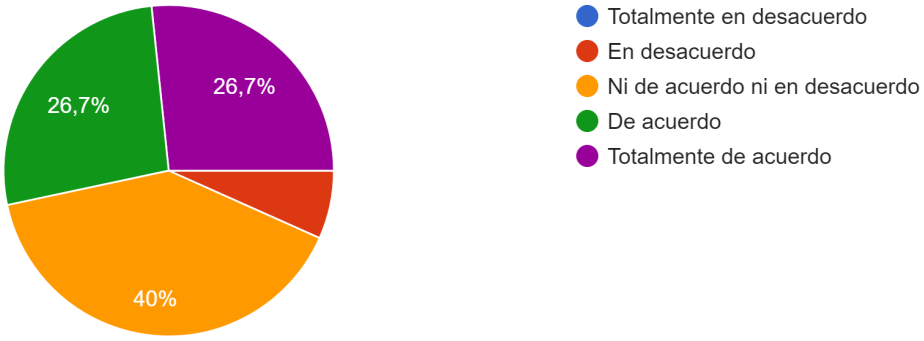


Figura 4.6: Resultado de la pregunta, ¿En general HAM Open Data Colombia es útil en su trabajo?. Fuente: Elaboración propia.

En la pregunta número 6 se muestra que un 40 % de las respuestas estan de acuerdo y otro 40 % ni de acuerdo ni en desacuerdo en que los metadatos ofrecidos por HAM son suficientes para escoger un buen conjunto de datos, como se muestra en la Figura 4.7.

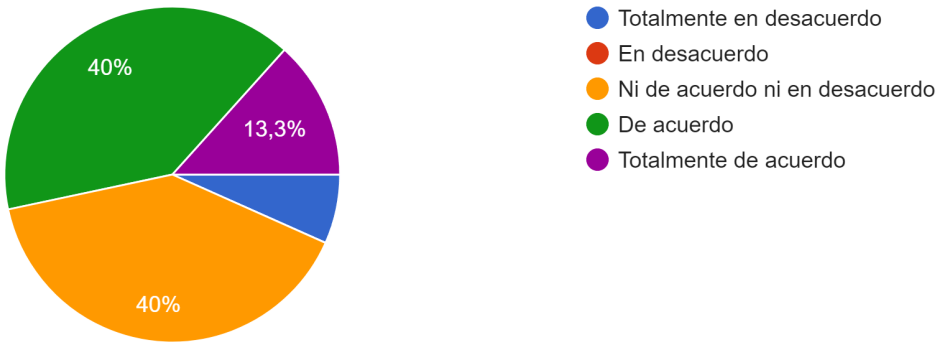


Figura 4.7: Resultado de la pregunta, ¿Los metadatos mostrados por HAM Open Data Colombia son suficientes para escoger un buen conjunto de datos?. Fuente: Elaboración propia.

Para la pregunta número 7 se observa un 46.7 % ni en desacuerdo ni de acuerdo y un 26.7 % de acuerdo en que recomienda los indicadores de HAM, como se muestra en la Figura 4.8.

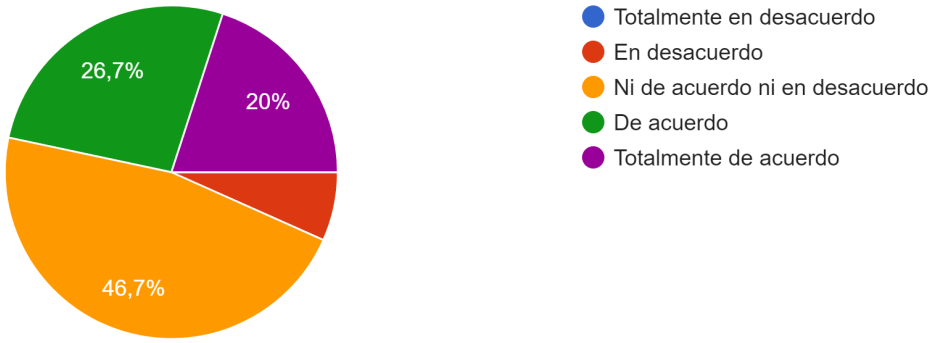


Figura 4.8: Resultado de la pregunta, ¿Recomendaría los indicadores propuestos por HAM Open Data Colombia?. Fuente: Elaboración propia.

Para la pregunta número 8 se muestra un 40 % de las respuestas ni deacuerdo ni en desacuerdo y un 33.3 % de acuerdo en que HAM es útil para los encuestados, como se ilustra en la Figura 4.9.

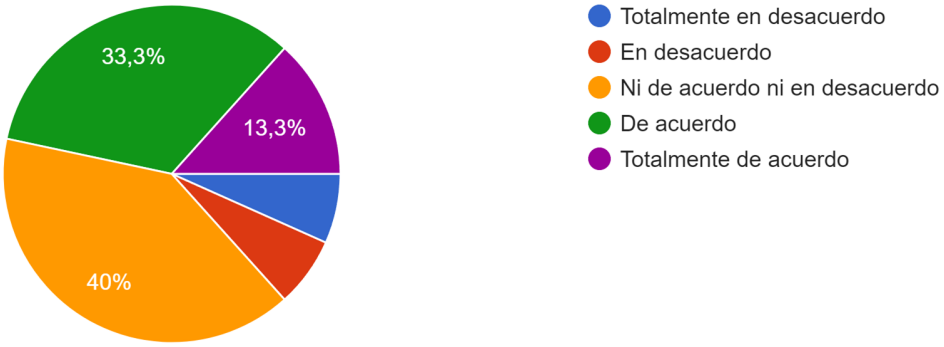


Figura 4.9: Resultado de la pregunta, ¿Con los indicadores mostrados, cree usted que HAM Open Data Colombia es útil?. Fuente: Elaboración propia.

Finalmente en la pregunta número 9 se observa un 53.3 % ni deacuerdo ni en desacuerdo y un 33.3 % de acuerdo en que se necesitan mas metricas para evaluar la utilidad de HAM, como se observa en la Figura 4.10.

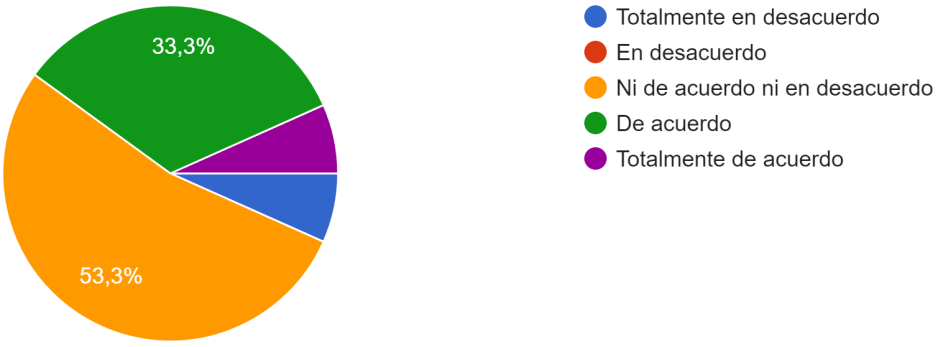


Figura 4.10: Resultado de la pregunta, ¿Considera usted que se necesitan más métricas para evaluar la utilidad de HAM Open Data Colombia?. Fuente: Elaboración propia.

4.1.4. Análisis general de tendencia

Como observación general de la evaluación del instrumento de medición, se continuó con un análisis de los resultados totales, para identificar cuál es la tendencia de dicha evaluación.

Para ello se totalizó en la siguiente tabla 4.4 el resultado de las frecuencias generales por pregunta.

Respuesta	**Peso**	**Valor total respuesta**	**Porcentaje total por respuesta**
Totalmente en desacuerdo	1	4	2.96 %
En desacuerdo	2	10	7.41 %
Ni en desacuerdo ni de acuerdo	3	55	40.74 %
De acuerdo	4	47	34.81 %
Totalmente de acuerdo	5	19	14.07 %

Tabla 4.4: Frecuencias de respuestas por pregunta porcentaje total. Fuente: Elaboración propia.

Con esta información podemos deducir que a pesar de que el 40.74 % de las respuestas manifiesta estar **Ni en desacuerdo ni de acuerdo** hay un porcentaje de 34.81 % y uno de 14.07 % que están de acuerdo y totalmente de acuerdo en que HAM es útil para el desarrollo de sus actividades, esto se ilustra de mejor manera en la siguiente Figura 4.11:

Figura 4.11: Resultado análisis general de tendencia. Fuente: Elaboración propia.

CAPÍTULO 5

Conclusiones

5.1. Concluciones y Trabajos Futuros

Segun el análisis realizado a través de la Herramienta para el Análisis de Medatados (HAM), se evidencia que la cantidad de metadatos presentes en la plataforma **datos.gov.co** es reducida, lo que dificulta establecer la utilidad de la plataforma a través de los metadatos.

Adicionalmente se observó que no hay controles sobre los datos que se ingresa a la plataforma, esto incrementa la dificultad de analizar dichos datos como un conjunto, debido a que los campos más representativos sobre el contenido de los datos, son su nombre y la descripción, siendo la descripción un campo de poco menos de 4.000 caracteres, que no tiene ningún formato, ni validación previa del contenido de estos campos, lo que permite que la información ingresada en estos no esté relacionada. Esto implica que si un usuario busca por nombre, cave la posibilidad de que la descripción no tenga relación con lo que busca en realidad, y en el peor de los casos el dataset o los datos tampoco.

La plataforma tampoco cuenta con validaciones sobre las fechas y los cambios que se hacen sobre los datos. No posee control sobre la continuidad de la información en el tiempo, lo que hace difícil para los usuarios que quieren usar los datos, encontrar información consistente, esto hace que la calidad en cuanto a la información que almacena la plataforma **datos.gov.co** baje.

Como resultado del estudio hecho por la Herramienta para el Análisis de Medatados (HAM) a la plataforma **datos.gov.co**, se puede concluir que la plataforma tiene mucho por mejorar en cuanto a metadatos disponibles, y que no está cumpliendo con el principio de generar nuevo conocimiento, el cual es fundamental para el movimiento de datos abiertos. Esto se lograria con la masificación del uso de los datos dispuesto en la misma, pero según lo análisado esto no es lo que sucede, esta conclución es consistente con los argumentos mostrados por (Abril Jiménez et al., 2017) donde indica que la plataforma **datos.gov.co** no cuenta con estándares adecuados en comparación con plataformas internacionales.

Este argumento se basa en la evaluación hecha a través del instrumento de medición, que a juicio de expertos, muestra que a pesar de que hay un 34.81 % que considera útil la plataforma, hay un 40.74 % que no tienen una posicion clara con respecto a la utilidad de la plataforma. Contrastando esto, aún se requiere de más información para poder concluir que la plataforma es útil.

Como trabajo futuro, se plantea realizar un análisis mas profundo, centrado en los metadatos de descripción y nombre, haciendo uso de procesadores de lenguaje natural, para encontrar relaciones que pueden

ayudar a complementar el análisis presentado en este documento.

También se propone como trabajo futuro, elaborar o diseñar un mecanismo para el control de los campos nombre y descripción de los datos, al momento de ser ingresados a la plataforma, con el objetivo de maximizar la relación entre estos campos y así optimizar la búsqueda.

Finalmente también se propone como trabajo futuro, elaborar o diseñar un control para que la plataforma tenga un mejor manejo de los cambios sobre los datos en el tiempo, ya que como se mencionó, los metadatos de fecha de última actualización y creación no cambian a pesar de que el dato se actualice.

Bibliografía

(1996). The data warehouse toolkit.

Abella, A., Ortiz-de Urbina-Criado, M., and De-Pablos-Heredero, C. (2014). Meloda, métrica para evaluar la reutilización de datos abiertos. *Profesional de la información*, 23(6):582–588.

Abril Jiménez, H. Y., Aguirre Santafé, F. M., and Montilla Garzón, Y. M. (2017). Guía de normalización de metadatos para datos abiertos.

Charter, O. D. (2015). International open data charter.

Council, N. R. (1985). *Methods for Designing Software to Fit Human Needs and Capabilities: Proceedings of the Workshop on Software Human Factors*. The National Academies Press, Washington, DC.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems*, 13:319–339.

Foley, É. and Guillemette, M. G. (2010). What is business intelligence? *International Journal of Business Intelligence Research*, 1(4):1–28.

Gobierno de Colombia (2014). `https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=56882`, 01 de Noviembre de 2022.

Gobierno de Colombia (2022a). `https://www.datos.gov.co/stories/s/Reportes-Portal-Nacional-Datos-Abiertos/pvyw-9yqs`, 01 de noviembre de 2022.

Gobierno de Colombia (2022b). `www.datos.gov.co`, 01 de noviembre de 2022.

Heise, D. R. and Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. *Sociological Methodology*, 2:104–129.

IBM (2022). `https://www.ibm.com/docs/es/ida/9.1.2?topic=modeling-dimensional`, 19 de noviembre de 2022.

Kassen, M. (2013). A promising phenomenon of open data: A case study of the chicago open data project. *Government Information Quarterly*, 30(4):508–513.

Melo, C. A. H. and Sanabria, J. S. G. (2020). Proposal for the evaluation of open data portals. *Revista Facultad de Ingenieria*, 29.

Moody, D., Kortink, M., Moody, D. L., and Kortink, M. A. R. (2000). From enterprise models to dimensional models: A methodology for data warehouse and data mart design.

MoralesVallejoPedro, UrosaSanzBelén, B. (2003). *Construcciones de escalas de actitudes tipo likert : una guía práctica*, page 23. La Muralla.

Murray-Rust, P. (2008). Open data in science.

Open Data HandBook (2022). https://opendatahandbook.org/, 01 de noviembre de 2022.

Open Knoledge Fundation (2022). https://www.opendefinition.org, 01 de noviembre de 2022.

Santos, V. and Belo, O. (2011). Slowly changing dimensions specification a relational algebra approach.

Tim Berners-Lee (2006). https://www.w3.org/DesignIssues/LinkedData.html, 27 de julio de 2006.

Tim-Berners-Lee (2022). https://www.w3.org/DesignIssues/LinkedData.html, 7 de noviembre de 2022.

vom Brocke, J., Hevner, A., and Maedche, A. (2020). *Introduction to Design Science Research*.