UNIVERSIDAD NACIONAL

Facultad de Ciencias Exactas y Naturales

ESCUELA DE INFORMÁTICA

"Ambiente de modelado de datos como servicio para facilitar el proceso de creación de modelos de datos en una organización"

Para optar al grado de Licenciado en Informática
con énfasis en Sistemas Web

Ing. Jefry Baltodano Zamora

Heredia, Costa Rica

Agradezco a mi tutor Cristopher Montero Jimenez, quien ha sido un excelente apoyo y guía en la elaboración de este proyecto. Sus amplios conocimientos en distintas áreas han sido de gran ayuda para completar este proyecto satisfactoriamente.

Agradezco a mi lector externo Ronald Meléndez Suarez, por su gran ayuda y excelentes consejos durante mi formación profesional y en la elaboración de este proyecto. Su impecable disposición y experiencia, han sido de mucha ayuda desde el inicio de mi carrera profesional.

Agradezco a mi lector interno Rodolfo Sánchez Sánchez, por su excelente disposición a colaborar y revisar el proyecto. Su área de especialización es de mucha relevancia para el proyecto y su aprobación es un gran privilegio para la conclusión del mismo. Asimismo, agradezco por los conocimientos que compartió conmigo durante mi desarrollo en la universidad.

TABLA DE CONTENIDOS

CAPÍTULO I: INTRODUCCIÓN *7*

1. Antecedentes 7

2. Planteamiento del problema 8

3. Justificación 9

4. Objetivos del Proyecto 13

4.1 Objetivo general *13*

4.2 Objetivos específicos *13*

CAPÍTULO II: MARCO TEÓRICO *15*

CAPÍTULO III: METODOLOGÍA *24*

1. Tipo de investigación 24

2. Población y muestra 24

3. Descripción de instrumentos 25

4. Procedimiento para analizar la información del diagnóstico 26

CAPÍTULO IV: PROPUESTA DE SOLUCIÓN *34*

1. Diagnóstico 34

2. Propuesta de solución 35

3. Validación de la propuesta 57

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES *62*

1. Conclusiones 62

2. Limitaciones 63

3. Trabajos futuros 64

REFERENCIAS *67*

Anexo 1 *75*

Anexo 2 *75*

Anexo 3 *75*

Anexo 4 *76*

Anexo 5 *78*

Anexo 6 *82*

Tablas

Tabla 1 17
Tabla 2 28
Tabla 3 31
Tabla 4 56

Imágenes

Imagen 1 35
Imagen 2 36
Imagen 3 36
Imagen 4 38
Imagen 5 38
Imagen 6 39
Imagen 7 39
Imagen 8 40
Imagen 9 41
Imagen 10 41
Imagen 11 42
Imagen 12 43
Imagen 13 44
Imagen 14 45
Imagen 15 45
Imagen 16 46
Imagen 17 46
Imagen 18 47
Imagen 19 47
Imagen 20 48
Imagen 21 48
Imagen 22 49
Imagen 23 49
Imagen 24 50
Imagen 25 51
Imagen 26 51
Imagen 27 52
Imagen 28 53
Imagen 29 53
Imagen 30 54
Imagen 31 55

CAPÍTULO I
INTRODUCCIÓN

CAPÍTULO I: INTRODUCCIÓN

1. Antecedentes

Con el continuo avance tecnológico y la integración cada vez mayor de sistemas de información en muchos de los procesos organizacionales, es natural que surja la necesidad de empezar a analizar los datos generados por estos sistemas con el fin de obtener un mayor valor de estos. De esta manera, se habilita la posibilidad de apoyar muchas de las decisiones que se toman en las organizaciones. Este análisis de datos, es conocido como inteligencia de negocios y se destaca por ser una fase crítica en la toma de decisiones en muchas organizaciones. A través de la investigación "El impacto de las herramientas de inteligencia de negocios en la toma de decisiones de los ejecutivos", Calzada & Abreu, (2009) se menciona que: "*las empresas dedican una importante parte de su tiempo y recursos en obtención, proceso, aplicación y proyección de información*" (p. 19). Dicho esto, esta información juega un papel decisivo en las organizaciones, convirtiéndose en su principal patrimonio.

Al ser la información un importante recurso para la toma de decisiones en las organizaciones, es importante contar con las herramientas adecuadas para obtener el mayor valor de los datos. Según Castro (2013) en su investigación Indicadores de gestión para la toma de decisiones basada en Inteligencia de Negocios , asegura que: "*es necesario que la organización cuente con mecanismos que le permitan tratar la información, de modo que se pueda integrar, unificar, interpretar y extraer lo más valioso*" (p. 87).

De igual manera, existen distintos tipos de sistemas de información que resultan convenientes para resolver diferentes problemas dentro de las organizaciones. Estos sistemas, ayudan a acelerar y a mejorar procesos. Basado en experiencias laborales, resulta fácil entender la importancia del software y los datos en las organizaciones. Aun así, no todas las organizaciones cuentan con acceso al mismo tipo de software, especialmente cuando este requiere más conocimiento técnico e infraestructura. A la hora de realizar análisis para la toma de decisiones, es conveniente contar con una herramienta de software que facilite el

proceso de modelado y visualización de datos, especialmente cuando la cantidad de datos crece cada día más.

Este proyecto se enfoca en el diseño de un ambiente de modelado de datos como servicio que brinde una alternativa adicional a la aplicación de inteligencia de negocios y a su vez, facilite el manejo de la información a través de modelos de datos que ayudan a contestar las distintas preguntas de negocio.

2. Planteamiento del problema

En la industria actual, no se ha logrado identificar un software que permita a los usuarios desarrollar sus propios modelos de datos sin importar que tan cambiantes o complejos puedan ser, y su vez realizar transformaciones en los datos sin tener que preocuparse como se modela o procesa la información en el backend. Es por esto, que surge la idea de brindar una alternativa al proceso de modelado de datos orientado a la aplicación de inteligencia de negocios, el cual es de mucha importancia para las organizaciones.

Al no contar con un ambiente de modelado de datos que brinde una abstracción al proceso de modelados de datos, las organizaciones se enfrentan a tener que realizar altas inversiones para instalar un ambiente de big data y análisis. Además, puede resultar muy complicado de mantener, generando problemas y retrasos constantes a la organización.

Resulta interesante seguir el siguiente ejemplo, el cual busca explicar la problemática y un caso de uso simple de cambios que pueden ocurrir en los modelos de datos y cómo estos pueden afectar el tiempo de respuesta de la organización y los cambios del mercado, perdiendo tiempo valioso que puede ser determinante para convertirse en líderes de un mercado potencial. [Ejemplo A] Se parte de la idea de tener datos crudos con columnas llamadas A, B y C, las cuales tienen los valores 2, 2 y 8 respectivamente. El encargado de estos datos necesita un modelo que siempre que los valores de A y B sean iguales, entonces tomar el valor de C y multiplicarlo por el valor de A. Generando un resultado final de: A, B y C con valores 2, 2 y 16 respectivamente. Como el encargado del modelo conoce la

necesidad de la organización, puede solicitar a un equipo de desarrollo una aplicación que permita realizar el cálculo anterior. Una vez terminada la aplicación, se habilita en un ambiente de producción, pero después el encargado del modelo se da cuenta que también necesita generar una columna adicional llamada D, si y sólo si el valor de C es mayor a 20; por lo tanto, deberá generar un nuevo requerimiento para agregarlo a la aplicación, lo cual se traduce en tiempo y costo, y que tampoco garantiza que no será necesario un nuevo cambio en el modelo pronto, que significa más gastos.

3. Justificación

Con el fin de proponer una solución a la problemática planteada, se busca brindar la posibilidad de hacer modelos de datos sin necesidad de cambiar requerimientos en el sistema ni preocuparse por el funcionamiento interno, es decir, un sistema que pueda brindar modelado de datos como un servicio. Basado en la idea, un usuario podría realizar la operación que se explica en el ejemplo anterior [Ejemplo A] y el día siguiente modificarla, eliminarla o incluso agregar nuevas operaciones sin tener que preocuparse del funcionamiento interno de la aplicación, solo de consumir el servicio.

Desde los inicios de cualquier compañía, se involucran con distintos procesos que requieren el uso y generación de datos, tanto físicos como digitales. Además, como mencionan los autores Delfín & Acosta (2016) , "*el desarrollo de las pequeñas y medianas empresas (Pymes) en el mercado global es una prioridad para el crecimiento económico de cada país".* (párr. 1)

De acuerdo con la cita anterior, resulta de alto valor apoyar estas empresas por medio de un ambiente de modelado de datos que los ayude a crecer de una mejor manera, ya que es evidente que los datos son parte fundamental de las compañías. Esta información, representa beneficios que tienen un alto impacto positivo en el crecimiento del negocio, y por consecuencia del país.

En la actualidad, los datos forman un aspecto fundamental en todas las organizaciones, siendo una herramienta clave para la toma de decisiones, encontrar nuevos clientes, seguir tendencias sociales, mejorar procesos de atención al cliente o mercadeo. De acuerdo con Cordero (2014) "*Según la firma de estudios de mercado, comScore, diariamente se crean en el mundo 2,5 quintillones de bytes como producto de tener 2,4 millones de usuarios de Internet y 1,5 millones de propietarios de smartphon*es" (párr. 2) En definitiva, los datos se han convertido en una herramienta valiosa que debe ser utilizada con inteligencia por las organizaciones.

Además, el ambiente de modelado de datos como servicio ayuda a responder de mejor manera a diferentes tipos de eventos, como lo pueden ser la creación de nuevos mercados, detectar nuevas tendencias o realizar una elección correcta del lanzamiento de sus productos para tener el mejor impacto en el negocio. Por lo tanto, tener un ambiente que permita modelar datos de una manera rápida y sencilla, habilita distintas oportunidades a las organizaciones para comenzar a realizar análisis de sus datos y a entender mejor cómo ser exitosos en su negocio. Un cambio en el modelo de datos puede afectar el tiempo de respuesta de una organización a un mercado potencial y cómo estos afectan la manera en cómo corre el negocio de la organización. Es por esto, que es importante contar con un ambiente de modelado de datos que brinda la flexibilidad a los usuarios de actualizar sus modelos de datos de una manera sencilla y así lograr llevar siempre a tiempo a mercados potenciales.

Un claro ejemplo de esto, es la manera en cómo las organizaciones utilizan datos para ofrecer productos y servicios de manera rápida e inteligente. Dicho esto, "*puede estar en un sitio web buscando zapatos, pero no realiza una compra. A medida que te mueves a otros sitios web y actividades en línea, de repente los anuncios de zapatos comienzan a aparecer dondequiera que vayas*". (Lewinter, 2016, párr.2). Todos estos datos pueden ser modelados y analizados para dirigir el negocio en la dirección correcta y en el momento adecuado, sin embargo, al no contar con el ambiente de modelado de datos, resulta difícil y requiere de tiempo extra para darle sentido a la información.

Este tipo de análisis se han venido realizando con el paso de los años, sin embargo, las tecnologías y la cantidad de información que se dispone en la actualidad ha cambiado exponencialmente, y las organizaciones deben cambiar la manera en cómo utilizan estos datos. Tal y como lo explica (Meyer, 2015) *"el software corporativo que recababa datos de todas las áreas de una organización en "almacenes de datos" (data warehouses) se analizaba periódicamente para descubrir las ideas más esclarecedoras respecto al rendimiento del negocio"* (párr. 1) .

Es fácil comprender que los datos han sido importantes para las organizaciones y lo seguirán siendo, es por esto que una optimización en la manera de modelar estos datos es de suma importancia, especialmente desde que el acceso a internet se incrementó exponencialmente como puede ser notado con la llegada de los teléfonos inteligentes. De acuerdo con Meyer(2015) *"los teléfonos móviles inteligentes se están haciendo universales, haciendo llegar el conocimiento a prácticamente todos los lugares".* (párr. 2) .

Además de hacer llegar conocimiento a nuevos lugares, los teléfonos inteligentes permiten generar mucha información que puede ser recolectada y analizada para conseguir indicadores fundamentales para la toma de decisiones en muchas organizaciones. Sin embargo, al no utilizar un ambiente de modelado de datos que facilite el proceso, resulta difícil para pequeñas y medianas organizaciones iniciar con la aplicación de inteligencia de negocios. Con un adecuado modelo de datos, ágil y cambiante, se puede mejorar el proceso de toma de decisiones basado en datos, lo cual ayuda a alcanzar clientes potenciales, entre otros beneficios.

Vale la pena agregar, que actualmente la información se mueve de forma acelerada, y se puede obtener desde cualquier lugar del mundo y a cualquier hora del día, convirtiéndose en un mar de información que procesar al alcance de muchas organizaciones. Estas compañías deben procesar grandes cantidades de datos, mayores a los que sus sistemas de información han sido desarrollados. También, se requiere una alta inversión para modernizar estos sistemas y procesar la información si no se utiliza un ambiente de modelado de datos que facilite el proceso.

Mucha de esta información se puede procesar de manera instantánea con el fin de generar indicadores valiosos para una organización. De la misma forma, "*las transacciones de tarjeta de crédito se producen instantáneamente, el análisis debería realizarse también en tiempo real*" (Mayer-Schönberger, Viktor; Cukier, Kenneth, 2013, p. 27).

De acuerdo con la cita anterior, en la actualidad existen herramientas que facilitan el movimiento de datos acercándose a tiempo real que pueden ser aprovechadas por las entidades para realizar análisis de grandes cantidades de datos de manera ágil y cambiante.

En la actualidad, se cuenta con dispositivos inteligentes y conectados que facilitan el análisis e integración de datos. Con la ayuda del ambiente de modelado de datos se puede gestionar modelos y consultar información de una manera sencilla, utilizando dispositivos conectados a internet. Con lo anterior, se evita preocuparse por complejas operaciones, ayudando a enfocarse en los datos y el negocio.

Para el desarrollo y pruebas de este proyecto se va a utilizar los datos de COVID-19, relacionados a Costa Rica, que son publicados por la unidad de investigación Observatorio del desarrollo de la Universidad de Costa Rica. Estos datos son de alta relevancia para el país y, a su vez, son muy útiles para analizar y obtener información extra de los mismos. También, se va a utilizar datos del INEC para el desarrollo de modelos de datos adicionales que permitan conocer la versatilidad del ambiente de modelado de datos.

4. Objetivos del Proyecto

4.1 Objetivo general

Implementar un ambiente de modelado de datos como servicio para facilitar el proceso de creación de nuevos modelos en una organización sin modificación de infraestructura o código de la aplicación.

4.2 Objetivos específicos

1. Investigar sistemas actuales de creación de modelos, consulta, almacenamiento y movimiento de datos, identificando el estado, tiempo y costo para conocer el estado actual del proceso de modelado de datos.
2. Diseñar un framework que permita el modelado y consulta de datos sin modificar infraestructura o código de la aplicación para brindar al usuario un ambiente de trabajo limpio y simple manteniendo el enfoque “as a service”.
3. Implementar un framework que permita consultar datos, crear nuevos modelos, transformarlos, almacenarlos y moverlos a través de la aplicación para solventar la problemática planteada.
4. Evaluar el ambiente de modelado de datos, mediante pruebas virtuales al framework desarrollado, para identificar los problemas que se resolvieron con el diseño del framework en términos de estado, tiempo y costo del proceso de modelado de datos.

CAPÍTULO II
MARCO TEÓRICO

CAPÍTULO II: MARCO TEÓRICO

Un dato se refiere a: "*información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho*" (Real Academia Española, s.f., definición 1). En muchas situaciones las organizaciones cuentan con grandes cantidades de datos que pueden ser difíciles de interpretar, es por esto, que es importante convertir estos datos en información útil para la organización.

Con el fin de convertir un conjunto de datos en información útil para la organización, es necesario aplicar un adecuado modelado de datos. El modelado de datos consiste en el proceso de describir los datos con el fin de darles sentido para luego poder ser consumidos. Así mismo, a través del modelado de datos, se busca estructurar los datos de una manera que puedan ser de utilidad y satisfacer un conjunto de requerimientos.

Según lo menciona Parms (2015):

> El modelado de datos es un proceso que lo ayudará a darle sentido a sus datos al definirlos y categorizarlos, y estableciendo definiciones y descriptores estándar para que todos los sistemas de información de su organización puedan consumir sus datos (párr. 4)

Adicionalmente, el proceso de modelado de datos, orientado a inteligencia de negocios, busca contestar distintas preguntas de negocio que tienen las organizaciones. De igual manera, estos modelos de datos ayudan a las organizaciones a obtener el mayor provecho de sus datos y así alcanzar sus objetivos estratégicos de una manera más eficiente.

Además del modelado de datos orientado a la inteligencia de negocios, existen otros tipos de modelado que ayudan a definir la estructura o esquema de una base de datos. Este tipo de modelado es comúnmente utilizado al diseñar una base de datos para un sistema de información por primera vez. Sin embargo, este tipo de modelado de datos no tiene el mismo

enfoque que un modelo de datos orientado a responder preguntas de negocio específicas, que es de lo que trata este proyecto.

Una vez que se cuenta con un adecuado modelo de datos para un requerimiento específico, es necesario aplicar análisis que permitan obtener el mayor conocimiento posible de la información generada por el modelo. Análisis se refiere al : "*estudio detallado de algo, especialmente de una obra o de un escrito*" (Real Academia Española, s.f., definición 2). Este análisis, permite tener las mejores perspectivas que ayudan a la toma de decisiones de las organizaciones.

Por otra parte, existen distintas formas de almacenar los datos, siendo las bases de datos relacionales y no relacionales (no SQL) las más populares. Una base de datos relacional es un tipo de base de datos que almacena datos relacionados entre sí, de una forma estructurada de acuerdo con un modelo relacional. Por otra parte, las bases de datos no relacionales no requieren tener un modelo relacional o una estructura preestablecida para almacenar los datos.

Las bases de datos relacionales son muy útiles cuando se conocen los datos que se van a guardar y son debidamente estructurados. Sin embargo, las bases de datos no relacionales permiten trabajar con datos que pueden cambiar con el tiempo y crecer horizontalmente.

Con el fin de conocer más acerca de estos tipos de base datos, se puede revisar la siguiente tabla, la cual busca mostrar la comparación de algunas de las características más relevantes para este proyecto.

Tabla 1

Tabla comparativa de bases de datos relacionales y no relaciones de las características más relevantes para este proyecto.

Característica	Base de datos relacionales	Base de datos no relacionales
Estructura de los datos	Filas de datos basado en tablas	Pares de llave y valor basado en documentos
Consultas	Es fácil realizar consultas complejas, sin la necesidad de tener tanto conocimiento técnico.	Es más difícil realizar consultas complejas ya que requiere más atención y trabajo para completarlas. No es tan amigable para el usuario.
Usabilidad	Al utilizar SQL, resulta más amigable al usuario.	Al utilizar normalmente JSON, no es tan amigable para el usuario y puede resultar muy confuso al inicio.
Escalabilidad	Permite escalar verticalmente, usualmente agregando hardware que permita más procesamiento. Usualmente, este hardware tiene un alto costo.	Permite escalar horizontalmente y distribuir la carga de trabajo en hardware de menor costo.

Fuente: Elaboración propia.

Debido a la naturaleza de los datos complejos que pueden tener las organizaciones, resulta conveniente utilizar bases de datos no relacionales en el ambiente de modelado de datos, ya que no se sabe con exactitud los cambios que los usuarios realizan en los modelos y tampoco se conocen los datos que se insertan. Este requerimiento, se alinea adecuadamente a las funcionalidades brindadas por las bases de datos no relacionales.

En la industria actual, existen sistemas que ayudan al proceso de gestión de grandes cantidades de datos. Estas soluciones, abstraen gran parte de la complejidad técnica que conlleva el mantenimiento de un ambiente productivo de esta naturaleza.

Amazon Redshift, es un servicio de almacén de datos en la nube que forma parte de Amazon Web Services. Durante una revisión de información sobre la herramienta, se logró identificar que el proceso de creación o generación del cluster, para almacenar los datos, es bastante sencillo, aunque requiere de un conocimiento técnico, especialmente a la hora de elegir la cantidad de CPUs, capacidad de memoria y de almacenamiento que se necesita. También, se identificó que la creación de los modelos debe ser hecha por los usuarios según los casos de uso que se tengan.

Una manera de crear los modelos de datos en Amazon Redshift, es utilizando SQL a través de algún gestor de base de datos. La conexión se puede realizar utilizando un jdbc driver y la cadena de conexión al cluster correspondiente. Una vez que la conexión está hecha, se puede hacer uso del almacén de datos utilizando SQL.

De igual manera, se realizó una investigación sobre BigQuery ya que es el almacén de datos multi nube de Google que permite realizar análisis a los datos. Se logró identificar puntos muy favorables como lo es la obtención de información en tiempo real y un acceso sencillo a la información. Así mismo, Google destaca la protección que ofrece a los datos y la alta disponibilidad que posee. BigQuery, en su versión preliminar, utiliza SQL estándar.

También se investigó sobre Snowflake, la cual es una compañía que se dedica al almacenamiento de datos en la nube. Esta empresa asegura, en su sitio web, que los científicos de datos que usan herramientas tradicionales gastan 80% del tiempo buscando y preparando datos en vez de usar ese tiempo para diseñar y construir modelos. También asegura que más del 87% de los proyectos de ciencia de datos nunca llegan a un estado productivo a pesar de haber realizado un gasto de millones de dólares en los procesos de diseño y construcción. Conociendo estos datos, ellos ofrecen un servicio seguro y escalable donde los científicos de datos hacen uso de una única fuente de datos que se mantiene siempre actualizada. Entre sus ventajas, mencionan que los científicos de datos descubren, al usar su solución, que tienen más tiempo para la exploración de nuevos modelos y prueba de nuevas

herramientas de aprendizaje automático o aprendizaje de máquina. Similar a otras alternativas, utilizar SQL para interactuar con los datos.

Después de la investigación a las alternativas de modelado de datos, no se logró identificar un software que permitiera crear modelos de la misma manera que lo hace el ambiente de modelado de datos propuesto en este documento.

Por otra parte, los sistemas de información ayudan a resolver problemas dentro de las organizaciones y en muchas ocasiones se necesita utilizar diversos sistemas en conjunto para satisfacer una necesidad u objetivo estratégico mayor. Es por esta razón, que el ambiente implementa distintos tipos de software que permiten resolver la problemática planteada, así como también procesos de instalación sencillos y automatizados para estas herramientas. De acuerdo con Campos, (2004).

> El éxito de Internet ha modificado algunos de los principios de la comunicación, ya que la rapidez de transmisión y la cobertura de las telecomunicaciones nos permiten acceder a la información en cualquier momento y nos ofrecen la oportunidad de empezar a romper las barreras del tiempo y de la geografía (p. 5).

Con base en la cita anterior, se puede notar que el internet es fundamental para habilitar oportunidades de colaboración entre empleados y entre distintas sedes de la organización, rompiendo así barreras de tiempo y geografía. Por esta razón, el ambiente de modelado de datos incluye una aplicación web que permite la gestión de los modelos, y a que vez, facilita el uso y mantenimiento ya que no es necesario instalar software adicional en las máquinas de los usuarios ni tampoco preocuparse por limitaciones entre sistemas operativos.

Cómo comprendemos del autor Ralph Jacobson *"El 90% de los datos en el mundo de hoy han sido creados en los últimos dos años".* (Jacobson, 2013). De acuerdo con lo anterior, podemos entender que los datos son muy importantes y que las compañías deben sacar el mayor provecho de estos para conseguir las mejores oportunidades en la industria

actual. Todos estos datos, al ser masivos, conforman big data y reflejan muchas oportunidades para aquellas organizaciones que los utilizan correctamente.

Al ser estos datos tan valiosos, resulta importante analizarlos todos, sin embargo, estos pueden ser difíciles de interpretar y darles sentido. De acuerdo con el autor Davenport (2013).

> Es importante recordar que el valor principal del big data proviene no de los datos en su forma cruda, sino de su procesamiento y análisis y de los insights, productos y servicios que emanan del análisis. Los cambios radicales en tecnologías y métodos de gestión del big data necesitan ir acompañados de cambios similarmente dramáticos en la forma en que los datos dan soporte a decisiones y a la innovación en productos/servicios (p. 30).

Según lo citado anteriormente, es importante modelar los datos y darles sentido de modo que se puedan analizar y obtener los mayores beneficios producto de este análisis, encontrando patrones ocultos y correlaciones. El ambiente, además de brindar software para crear modelos de datos, provee herramientas para el análisis y la consulta de estos datos, de modo que el usuario puede sacar el mayor provecho, dándole un mejor uso a la información.

Gracias al análisis realizado a los datos, se habilitan oportunidades de innovación en la organización, las cuales permiten obtener ventajas competitivas con respecto a otras compañías. Por consiguiente, es fácil comprender que no es lo mismo para una organización seguir a sus competidores que ser el primero en ofrecer un producto o servicio. Por esta razón, el ambiente ofrece herramientas big data y de análisis de datos de una manera sencilla, donde las organizaciones pueden enfocarse en la inteligencia de negocios y no en las dificultades técnicas de un ambiente de big data y analytics.

Como parte de los análisis, es importante comprender que no todas las organizaciones necesitan realizar los mismos tipos de estudios en todo momento, es decir, en un momento particular puede ser valioso realizar análisis predictivo, pero después resulta más valioso

realizar análisis en tiempo real. Por lo cual, el framework ofrece distintas herramientas de análisis de datos de modo que el usuario utilice la más conveniente, así como también se deja abierta la oportunidad de que el usuario utilice algún software de análisis o consulta de datos de su preferencia.

Además de innovar, es importante contar con herramientas que permitan asegurar que se cuentan con los datos más actualizados a la hora de aplicar la inteligencia de negocios y tomar decisiones. Tal y como lo expresa Olivera (2016), "*El acceso a información útil ha sido un reto que al día de hoy muchas organizaciones no han podido resolver y siguen operando y tomando decisiones con información limitada, imprecisa y desactualizada*". (párr. 2).

Con base en la cita anterior, se puede comprender cómo esta limitación ha estado afectando la manera en cómo las organizaciones realizan análisis ya que no cuenta con los datos más recientes o precisos. Es por esto, que el framework hace uso de herramientas que permiten manipular grandes cantidades de datos de manera eficiente, lo cual ayuda a tener certeza que se cuentan con los datos más actualizados a la hora de realizar los análisis.

Existen ocasiones donde las organizaciones se enfrentan a incrementos en la cantidad de datos que reciben y deben procesar, como lo pueden ser las fechas festivas donde las personas suelen comprar más. Debido a esto, se debe contar con un sistema que sea escalable, logrando aumentar la capacidad de trabajo cuando sea necesario y así evitar latencia en los datos de la organización, todo esto, sin comprometer el correcto funcionamiento del sistema. Por esta razón, el ambiente cuenta con la capacidad de ejecutarse en una infraestructura de alta disponibilidad y escalabilidad.

Por otra parte, las organizaciones necesitan proteger su información como un activo de alto valor, ya que, si llega a la competencia, pueden perder toda la ventaja y, en algunos casos, el trabajo realizado. Además, se puede poner en riesgo información confidencial de clientes, lo cual produce una pérdida de confianza a la organización. Al ser la seguridad parte fundamental en los datos y en las organizaciones, el ambiente de modelado de datos hace uso

de tecnologías que son altamente utilizadas en la industria actual y que cuentan con soporte, de modo que reciben actualizaciones en caso de que se detecte alguna vulnerabilidad.

CAPÍTULO III
METODOLOGÍA

CAPÍTULO III: METODOLOGÍA

1. Tipo de investigación

1.1 Exploratoria

Así mismo, según como lo indica Landeau, (2007) se entiende por investigación exploratoria como: *"(...) un tema poco explorado o que no ha sido abordado, con el objeto de obtener conocimiento con respecto a la materia objeto de investigación*" (p. 56). Bajo este efecto, esta investigación es de tipo exploratoria, ya que es aplicada al diseño y desarrollo de un software que permite resolver una problemática para la cual no se ha logrado identificar ningún otro software que lo resuelva de la misma manera que lo hace el ambiente de modelado de datos propuesto en este documento.

1.2 Descriptiva

Para el presente trabajo, se recurre a la aplicación de la investigación descriptiva, por el hecho que se obtiene información valiosa de las características del modelado de datos y de la importancia de los datos en las organizaciones. Según expresa Bernal, (2010) "*(...)es la capacidad para seleccionar las características fundamentales del objeto de estudio y su descripción detallada de las partes, categorías o clases de ese objeto"* (p. 113).

2. Población y muestra

2.1 Poblacion

Seguidamente, se detalla la población de estudio. Se define población como: *"Un conjunto finito o infinito de elementos con características comunes para los cuales serán extensivas las conclusiones de la investigación. Esta queda determinada por el problema y por los objetivos del estudio"* (Arias, 2006, p. 81).

La población de interés para la presente investigación, son las organizaciones que cuentan con procesos de modelado de datos o están interesadas en iniciar estos procesos, pero

que no cuentan con la capacidad de investigación para implementar un ambiente de modelado de datos como servicio propio.

2.2 Muestra

Se puede definir muestra como : "*un subconjunto representativo y finito que se extrae de la población accesible*" (Arias, 2006, p. 83). La muestra está conformada por los colegas informáticos que laboran en organizaciones que pueden hacer uso del ambiente de modelado de datos sin tener que preocuparse por complejas operaciones, sino sólo de consumir el servicio ofrecido por el ambiente.

3. Descripción de instrumentos

De acuerdo con el autor, *"Los instrumentos en la investigación documental son técnicas que se han empleado para la recolección de datos de las diversas fuentes consultadas y que facilitan el manejo de la información*" (Ocegueda, 2004, p. 122).

Bajo este contexto, los instrumentos son parte esencial del proyecto ya que permiten obtener y recopilar información muy valiosa para la investigación. Esta investigación, hace uso de distintos instrumentos que se especifican a continuación.

Una encuesta se refiere a un "*conjunto de preguntas tipificadas dirigidas a una muestra representativa de grupos sociales, para averiguar estados de opinión o conocer otras cuestiones que les afectan*" (Real Academia Española, s.f., definición 1).

Inicialmente, se utiliza como instrumento una encuesta que permite conocer más acerca de la importancia de los datos, soluciones similares y datos relevantes para el país. Según afirma Zapata (2006) *"el investigador para presenciar directamente el fenómeno que estudia , sin actuar sobre él esto es, sin modificarlo o realizar cualquier tipo de operación que permita manipular"* (p. 110).

Con base en la cita anterior, se realiza una observación de tecnologías a través de una revisión de literatura que permite tener una mejor percepción de las herramientas que pueden ser utilizadas en el diseño y desarrollo del ambiente de modelado de datos.

Bajo este mismo marco, según la Real Academia Española, (s.f.) El concepto "consultar" se define como: "*Examinar, tratar un asunto con una o varias personas"* (definición 1). Dicho esto, para el presente proyecto se tiene una consultoría con un profesional de base de datos y diseño de software que facilita información esencial para el diseño del ambiente de modelado de datos.

4. Procedimiento para analizar la información del diagnóstico

El Project Management Institute (2008) define la estimación de costos como :"*el proceso que consiste en desarrollar una aproximación de los recursos necesarios (humanos y materiales) para completar las actividades de un proyecto*" (p. 52).

Dicho esto, inicialmente, se realizó una estimación de costos para una organización interesada en adoptar este ambiente de modelado de datos. Dicha estimación, está orientada a una solución básica para una empresa mediana de aproximadamente 100 empleados, usando una transferencia base mensual de 2 terabytes. Esta estimación de costos, está reflejada en el anexo 1.

Adicionalmente, se procedió a aplicar una encuesta. Esta encuesta, se puede encontrar en el anexo 3. La aplicación de una encuesta es importante para el diseño del ambiente de modelado de datos ya que permite comprender mejor los siguientes puntos:

- La importancia de los datos en las empresas.
- Identificar soluciones similares para modelado de datos.
- Datos de importancia para el país que pueden ser utilizados para realizar las pruebas.
- La utilización que se le daría a una base de datos relacional versus una base de datos no relacional.

Dicha encuesta, fue contestada por profesionales del área de la informática y de administración de empresas que trabajan continuamente con bases de datos que contienen información de alta importancia para distintos procesos empresariales. Después de la aplicación de la encuesta, se observa que un 100% de las respuestas indican que los datos son muy valiosos para las empresas. Además, la encuesta demuestra que el 100% de las personas encuestadas creen que un adecuado análisis de los datos es estrictamente necesario para el éxito de una empresa a largo plazo.

A través de la encuesta, se logra identificar algunas herramientas que pueden contribuir al desarrollo de un modelo de datos. Sin embargo, después de una revisión de las tecnologías, se logra comprender que estas herramientas no ofrecen una experiencia de modelado de datos como servicio, lo cual es el objetivo principal del ambiente de modelado de datos descrito en este documento.

Adicionalmente, la encuesta muestra que un 60% de los profesionales encuestados utilizarían una base de datos relacional para almacenar datos que pueden cambiar con el tiempo y que necesitan estar bien estructurados, mientras que el 40% utiliza una base de datos no relacional para el mismo propósito. Debido a esto, se procede a realizar una observación de estas tecnologías a través de una revisión de literatura en revistas web y documentación oficial de las herramientas.

Esta observación, se complementa con dos conversaciones de consultoría que permiten obtener información de suma importancia para el diseño del ambiente de modelado de datos. Inicialmente, se tiene una consultoría con un profesional en bases de datos con muchos años de experiencia en el diseño y desarrollo de sistemas de información y ex profesor de la Escuela de Informática de la Universidad Nacional de Costa Rica (UNA). Posteriormente, se tiene una consultoría con un profesional en base de datos con amplia experiencia en el diseño y desarrollo de sistemas de alta disponibilidad que permite obtener una visión más clara sobre aspectos indispensables a considerar durante el diseño del ambiente de modelado de datos.

Con la información recopilada en la observación y la consultoría aplicada, se decide utilizar una base de datos no relacional. La base de datos elegida es MongoDB ya que permite adaptarse fácilmente a todos los cambios que un modelo de datos pueda tener con el tiempo. Esta versatilidad, es muy beneficiosa en el diseño del ambiente de modelado de datos ya que facilita el almacenamiento de la información de acuerdo con la estructura actual del modelo sin tener que preocuparse por el esquema específico en la base de datos.

También, cuando es necesario aplicar un cambio a la estructura de un modelo de datos, no se deben realizar costosas operaciones para actualizar el esquema de la base de datos. Tampoco se debe actualizar los datos en sí, esto ayuda a potencialmente reducir la cantidad de tiempo necesaria para que el sistema pueda aplicar completamente un cambio en un modelo de datos.

Al entender, a través de la encuesta realizada, que los datos son tan esenciales para las empresas, se propone utilizar datos a nivel general sobre empleo generados por el Instituto Nacional de Estadística y Censos de Costa Rica (INEC) con el fin de probar el ambiente de modelado de datos desarrollado, utilizando datos de mucha importancia para el país. También, se va a utilizar datos sobre COVID-19 liberados por la unidad de investigación Observatorio del desarrollo de la Universidad de Costa Rica que permiten conocer la versatilidad del ambiente de modelado de datos.

Con el fin de establecer las herramientas que mejor se adaptan al desarrollo del ambiente de modelado de datos, se hace una revisión de literatura de las principales tecnologías que se utilizan actualmente para desarrollo de sistemas web y así poder tener un mejor criterio de estas, entendiendo cuál es su propósito y su caso de uso. Todo esto, con el objetivo de seleccionar cursos virtuales que ayuden a comprender mejor cómo utilizarlas y, así, definir posibles ventajas y desventajas a la hora de incluirlas en el ambiente de modelado de datos.

Durante la revisión de estas tecnologías, se conversa con distintos profesionales de informática y del área de tecnologías de datos para conocer acerca de las

tecnologías que deberían tener prioridad. De esta manera, se puede tener una guía más clara a la hora de invertir tiempo aprendiendo nuevas tecnologías para análisis y procesamiento de grandes cantidades de datos.

Como resultado de estas conversaciones y de la revisión de las tecnologías, se logra establecer una lista con distintos cursos virtuales que son de gran ayuda para el desarrollo profesional. Además, ayudan a comprender mejor cómo utilizar estas tecnologías para obtener el mejor beneficio al momento de diseñar y desarrollar el ambiente de modelado de datos.

Adicionalmente, sirven como fuente de información sobre aspectos importantes a considerar en temas de seguridad y patrones de diseño. Estos cursos, se encuentran publicados en los sitios educativos llamados Pluralsight.com y linkedin.com/learning.

Durante varias semanas, se ha estado estudiando estos cursos virtuales los cuales se basan en las siguientes tecnologías:

Tabla 2

Tabla de cursos virtuales.

Nombre del curso	Objetivo	Fuente
Angular: Getting Started	Reforzar las bases de Angular	Pluralsight.com
Getting Started with Docker	Conocer mejor acerca de los contenedores y como poder ser de gran ayuda a la hora de realizar el framework.	Pluralsight.com
Docker Deep Dive	Conocer más a fondo acerca de contenedores usando Docker.	Pluralsight.com

Docker and Kubernetes: The Big Picture	Tener mejor claridad en cómo gestionar múltiples contenedores y poder entender las ventajas y desventajas al momento de usarlas en el ambiente de modelado de datos.	Pluralsight.com
Managing Docker Images	Comprender cómo se construyen imágenes en Docker para así poder obtener el mejor provecho al momento de crear imágenes propias para el ambiente de modelado de datos.	Pluralsight.com
Getting Started with Kubernetes	Conocer más acerca de Kubernetes y de las ventajas que ofrece a la hora de utilizarlo en el ambiente a desarrollar.	Pluralsight.com
Learning Hadoop	Conocer mejor Hadoop para entender de una mejor manera su utilidad en el desarrollo del ambiente.	linkedin.com/learning
Learning the Elastic Stack	Conocer otras opciones de herramientas que permiten ingresar y consultar grandes cantidades de datos.	linkedin.com/learning
Kafka Essential Training	Conocer mejor una herramienta para transferencia de grandes cantidades de datos.	linkedin.com/learning
Designing and Implementing SQL Server Database Indexes	Entender mejor el uso de los índices en SQLServer para definir ventajas y desventajas de esta tecnología a la hora de incluirla en el ambiente de modelado de datos.	Pluralsight.com

Fuente: Elaboración propia.

Además de los cursos virtuales, se realizó una revisión de literatura que permitió conocer más acerca del proceso de modelado de datos y de las tecnologías que podían ser utilizadas en el proyecto. La revisión de literatura incluyó artículos, revistas, investigaciones previas y publicaciones en distintos sitios web.

También, se hace uso de la experiencia propia. Gracias a la experiencia laboral adquirida en el desarrollo de sistemas orientados al modelado de datos y a la continua involucración en el diseño y desarrollo de sistemas en la nube, se logra obtener conocimientos en distintas áreas de base de datos, software e infraestructura. Estos conocimientos adquiridos, ayudan mucho a establecer una propuesta de diseño del ambiente de modelado de datos que resulte favorable y factible para su aplicación en las organizaciones, resolviendo la problemática planteada. Al complementar los conocimientos adquiridos en la experiencia laboral con la revisión de literatura, se logra generar un mejor diseño del ambiente de modelado de datos y favorece enormemente el desarrollo de este.

Con base en la experiencia propia y la revisión de literatura, se logra continuar con el diseño del ambiente de modelado de datos. Se consideró utilizar tecnologías que permitieran desarrollar un ambiente escalable y de alta disponibilidad, así como también que permita interactuar con grandes cantidades de datos. Para la interfaz de usuario, se decidió utilizar Angular, ya que es un framework que resulta fácil de usar y permite desarrollar componentes que pueden ser utilizados a través de toda la aplicación de una manera sencilla, mantenible y fácil de probar con métodos automatizados.

Un punto que fue determinante al escoger Angular como framework para la interfaz de usuario, es que una aplicación escrita en Angular resulta fácil de ejecutar en un contenedor y esto representa una gran ventaja a la hora de impulsar la creación de un ambiente escalable y de alta disponibilidad debido a que los contenedores brindan distintas ventajas para este tipo de arquitecturas de software.

Para el desarrollo del software de lado del servidor y el API, se decidió utilizar Python ya que es un lenguaje fácil de usar y que se adapta muy bien los requerimientos del ambiente

de modelado de datos. También, al ser un lenguaje fácil de usar, resulta sencillo darle mantenimiento y desarrollar pruebas unitarias para asegurar que el código sigue funcionando como se requiere.

Otro punto que se tomó en cuenta para elegir Python para el desarrollo del servidor y del API, es que Python es conocido por muchos analistas de datos y de hecho, la interfaz de usuario brinda una opción para crear modelos de datos utilizando Python. De este modo, resulta fácil para el código del servidor interactuar con el código Python proveniente de la interfaz de usuario de manera ágil y segura.

Adicionalmente, se llega a la resolución de que el proyecto es técnicamente factible para las organizaciones dado que el ambiente puede ser ejecutado en la nube pública, lo cual reduce drásticamente muchas de las complejidades de instalación y mantenimiento de la infraestructura. También, los servicios del ambiente pueden ser accedidos desde dispositivos conectados a internet, lo cual es beneficioso y factible para las organizaciones ya que permite interactuar con el ambiente desde distintas ubicaciones. Así mismo, los costos de infraestructura y mantenimiento son menores en comparación a otras alternativas en el mercado. Además, los recursos de personal técnico para administrar el ambiente son muy pocos ya que mucho del soporte está incluido en la adquisición de los servicios de infraestructura en la nube pública. Todo esto, considerando su aplicación en una organización mediana y los aspectos técnicos mostrados en la siguiente tabla:

Tabla 3

Tabla de factibilidad técnica.

Producto o servicio	Consideraciones
Personal técnico	Un recurso encargado de administrar y mantener el ambiente
Máquina virtual en Azure	2 CPUs, 4GB memory, 32GB SSD
Instancias de contenedores	4 CPUs, 16GB memory
Almacenamiento en base de datos	2 TB
Licencias	Se utiliza software libre o con licencias sin costo asociado

Fuente: Elaboración propia.

CAPÍTULO IV
PROPUESTA DE SOLUCIÓN

CAPÍTULO IV: PROPUESTA DE SOLUCIÓN

1. Diagnóstico

Según lo indica (González, 2020) el diagnóstico :

> Es un método de estudio mediante el cual se logra conocer lo que ocurre en una situación específica. Es decir, se trata del análisis de una serie de sucesos con el objetivo de identificar los factores que promovieron la aparición de un fenómeno. (párr. 1).

Bajo este contexto, se puede inicialmente mencionar que es un hecho que los datos son parte fundamental en todas las organizaciones, además, es claro que son una herramienta clave para la toma de decisiones, detectar tendencias sociales, descubrir nuevos clientes y mejorar procesos de atención al cliente o mercadeo. De igual manera, es claro que, con la llegada del internet y los dispositivos inteligentes, los datos han crecido exponencialmente. Debido a esto, cada vez resulta más difícil, para los sistemas de información tradicionales, procesar e interpretar estas grandes cantidades de datos.

Otro aspecto importante, es que el proceso de análisis e interpretación de datos resulta costoso e ineficiente para las organizaciones, si no se cuentan con las herramientas correctas. Estas herramientas, son comúnmente utilizadas por grandes organizaciones, las cuales cuentan con altos recursos económicos y tienen alta capacidad de investigación.

Por otra parte, existe un desconocimiento acerca de estas áreas informáticas, en aquellas organizaciones que inician sus procesos, o bien, no cuentan con recursos para investigarlas. Lo anterior ocurre, debido a que, no se ha logrado identificar un framework, en la industria actual, que facilite el proceso de adopción de estas tecnologías, desde un punto de vista de modelado de datos para responder a preguntas de negocio.

Al no contar con un framework o ambiente que ayude a adoptar procesos de inteligencia de negocios y analítica, resulta aún más difícil para las pequeñas y medianas

organizaciones iniciar o expandir sus procesos en estas áreas. Esto ocurre, debido a que investigar acerca de todas estas tecnologías, resulta costoso y no permite a la organización enfocarse en sus clientes y negocio principal.

2. Propuesta de solución

Se propone un ambiente de modelado de datos, que facilita la incorporación o mejora de procesos de inteligencia de negocios, donde las organizaciones pueden sacar mayor ventaja de sus datos y, de esta manera, responder de la mejor manera a los distintos tipos de eventos que se generan cada día. Con la ayuda de este ambiente, se puede detectar de una manera más eficaz y eficiente, distintas tendencias y posibles nuevos mercados, así como también, facilita la toma de decisiones en distintos escenarios, como lo puede ser, a la hora de lanzar un producto o iniciar un nuevo negocio.

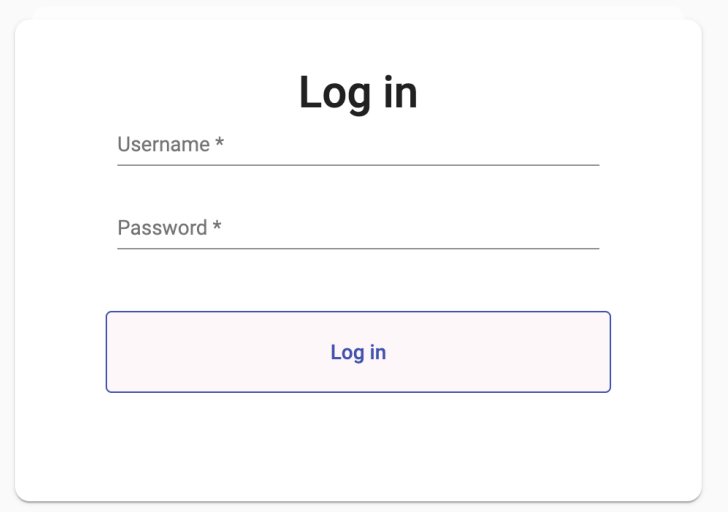
De la misma manera, el ambiente de modelado de datos facilita a las organizaciones a entender mejor cómo ser exitosos, a través de modelos de datos que responden a las distintas preguntas de negocio que tiene la organización. Con estas ventajas, las organizaciones pueden iniciar rápidamente a obtener el mayor provecho de sus datos, sin tener que invertir en costosos procesos de investigación para adoptar estas tecnologías.

Por otra parte, en la actualidad, se cuenta con dispositivos inteligentes y conectados en casi todas partes. Estos dispositivos, resultan ser una herramienta muy útil a la hora de acceder a sitios web o software en la nube. Es por esto, que el ambiente permite a los usuarios acceder a los modelos de datos desde estos dispositivos, y así, brindar agilidad y ayudar a la organización a ir más rápido y mantenerse enfocado en su negocio principal.

El ambiente de modelado de datos, cuenta con una pantalla para iniciar sesión. Esta pantalla, asume que el usuario ya está registrado en el sistema al ser éste un usuario de la organización. Es muy recomendable autenticar al usuario a través de un inicio de sesión único, sin embargo esta opción no fue habilitada en esta propuesta. Esta forma de autenticación, se sugiere en la sección “trabajos futuros” del capítulo V de este documento.

Imagen 1

Imagen de inicio de sesión.



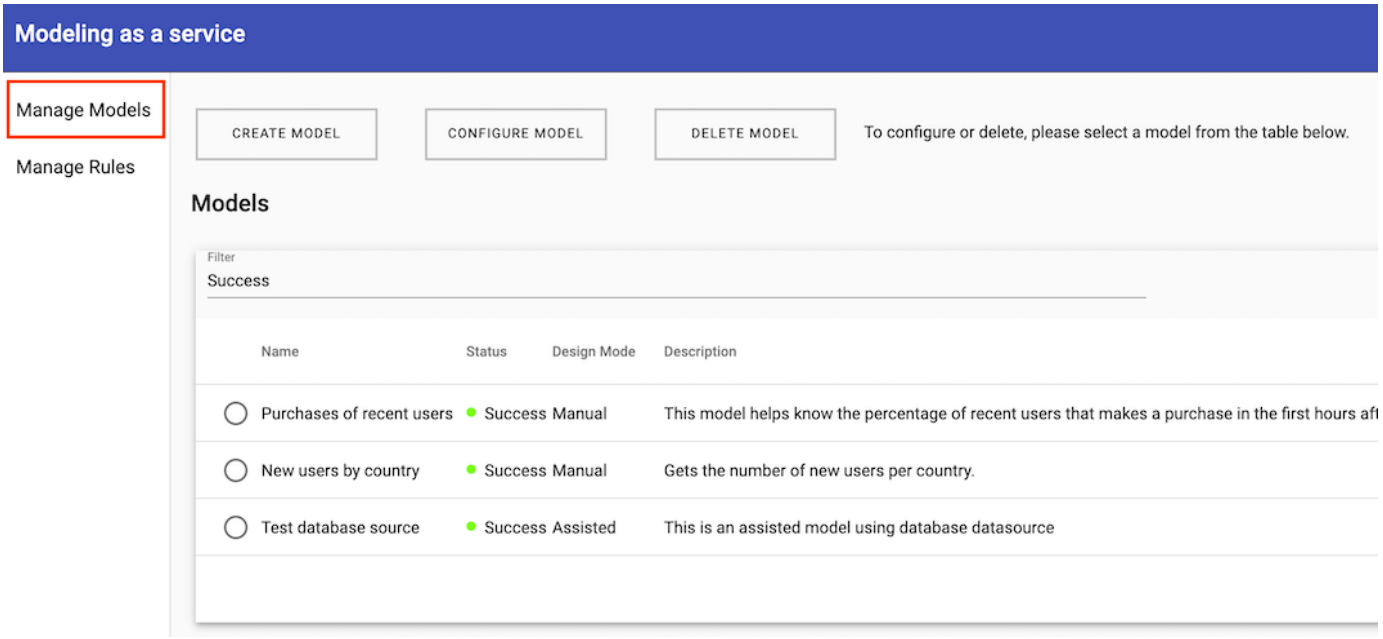
Fuente: Elaboración propia.

Adicionalmente, el ambiente cuenta con una interfaz web que permite a los usuarios interactuar con los modelos. En lo que a esta propuesta refiere, un modelo es una representación de un conjunto de configuraciones y acciones que se deben aplicar a cierto conjunto de datos con el fin de extraer información valiosa, generalmente para ayudar a una organización a contestar preguntas de negocio.

En la siguiente captura de pantalla, se puede observar la página principal para administrar los modelos de datos, la cual permite crear, configurar y borrar modelos, así como también, conocer información relevante de los modelos de datos a través de una tabla llamada “Models”.

Imagen 2

Imagen de vista para gestionar modelos.

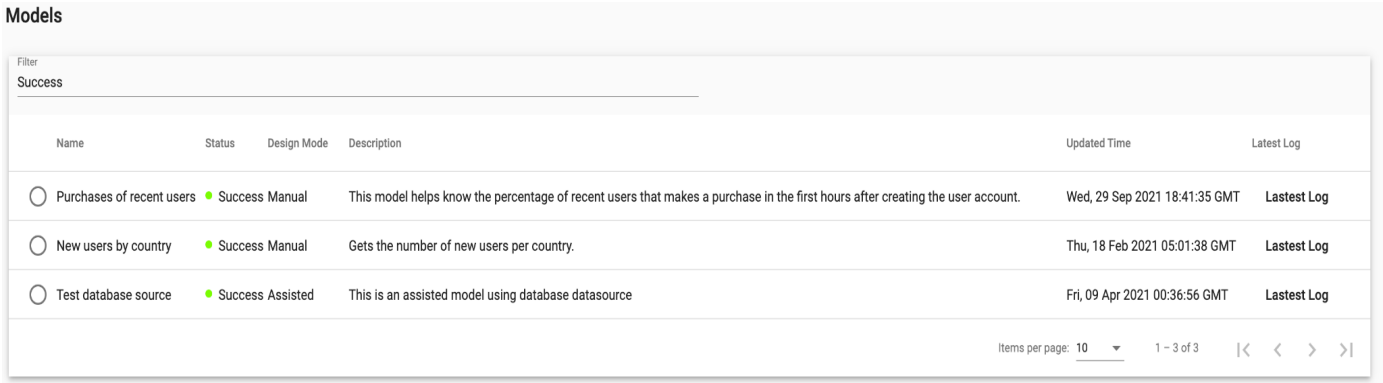


Fuente: Elaboración propia.

La tabla "Models", que se observa en la siguiente captura, muestra distinta información como lo es: nombre, estado, modo de diseño, descripción, última actualización y último registro.

Imagen 3

Imagen de tabla de modelos.



Fuente: Elaboración propia.

Esta tabla, cuenta con un filtro en la parte superior, el cual está habilitado para encontrar el mejor resultado utilizando todas las columnas de la tabla. También, la tabla cuenta con controles en la parte inferior que ayudan al usuario a manipular la información que se despliega en la tabla en cierto momento. Así mismo, esta tabla ayuda al usuario a seleccionar el modelo que desea gestionar a través de un botón de radio mostrado en la primera columna de la tabla.

El propósito de cada columna se detalla a continuación, a excepción de la primer columna que ya es de conocimiento del lector:

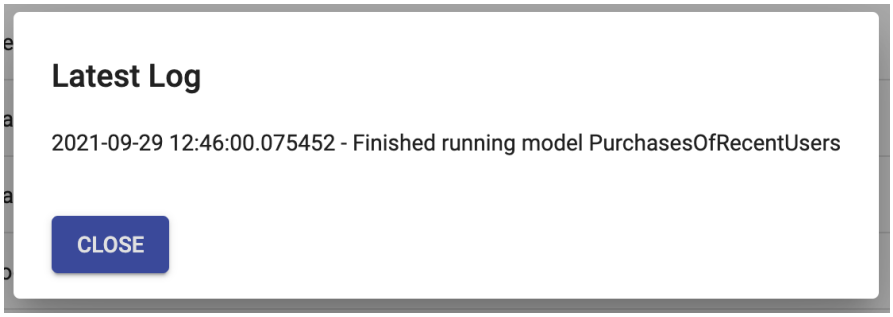
- **Name:** Es el nombre del modelo de datos asignado por el usuario durante la creación del modelo.
- **Status:** Es el estado actual del modelo. Los estados posibles son: scheduled, success, failed.
    - **Scheduled:** Este estado representa que el modelo de datos es nuevo o ha sido editado y se encuentra agendado para ser ejecutado próximamente, de acuerdo a su configuración.
    - **Success:** Este estado, indica que el modelo fue ejecutado exitosamente y que los datos están actualizados.
    - **Failed:** Este estado, indica que el modelo de datos fue ejecutado, pero se produjo un error al ejecutar sus acciones y los datos no fueron actualizados.
- **Design Mode:** Índice el modo de diseño utilizado para crear el modelo. Los modos de diseño disponibles son asistido y manual. Estos modos de diseño se detallarán más adelante en esta propuesta.
- **Description:** Muestra la descripción del modelo de datos que fue detallada por el usuario. Esta columna sirve como una ayuda a los usuarios para comprender o recordar el propósito de cada modelo de datos.
- **Updated Time:** Indica la última vez que el modelo de datos fue actualizado en términos de su configuración. Es importante destacar, que esta columna no indica el estado de actualización de los datos.
- **Latest log:** Esta columna muestra un botón que permite a los usuarios observar el último registro generado por el modelo de datos. Esta funcionalidad, es especialmente

útil cuando se quiere saber la razón del fallo en un modelo de datos. Seguidamente, se muestran unos ejemplos de esta funcionalidad.

- Último registro cuando el estado del modelo es “success”:

  Imagen 4

  Imagen de último registro con estado “success”.

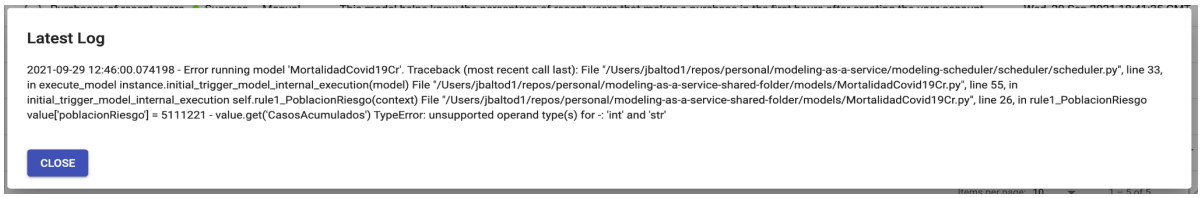
  

  Fuente: Elaboración propia.

- Último registro cuando el estado del modelo es “failed”:

  Imagen 5

  Imagen de último registro con estado “failed”.

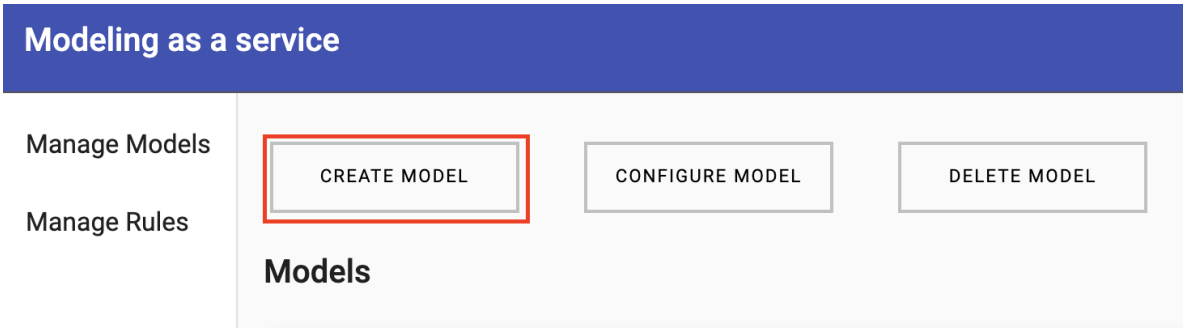
  Fuente: Elaboración propia.

- Último registro cuando el estado del modelo es “scheduled”: En este caso, existen dos escenarios:
  i. Si el modelo es nuevo, se muestra vacío.
  ii. Si el modelo ya existe y fue editado, se muestra el último registro de su estado guardado antes de ser editado.

El botón “create model”, como su nombre sugiere en inglés, permite crear un nuevo modelo.

Imagen 6

Imagen de botón para crear modelos.

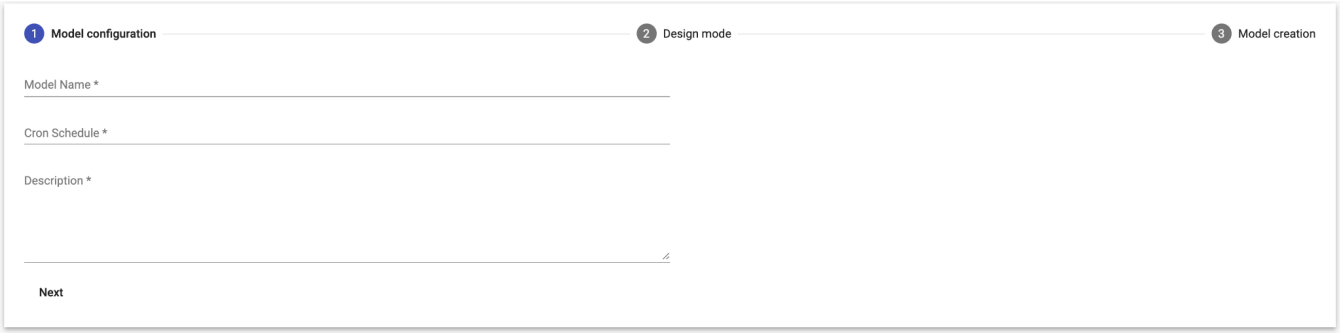


Fuente: Elaboración propia.

Seguidamente, se puede observar la pantalla mostrada al utilizar el botón “create model”:

Imagen 7

Imagen de pantalla crear modelos.



Fuente: Elaboración propia.

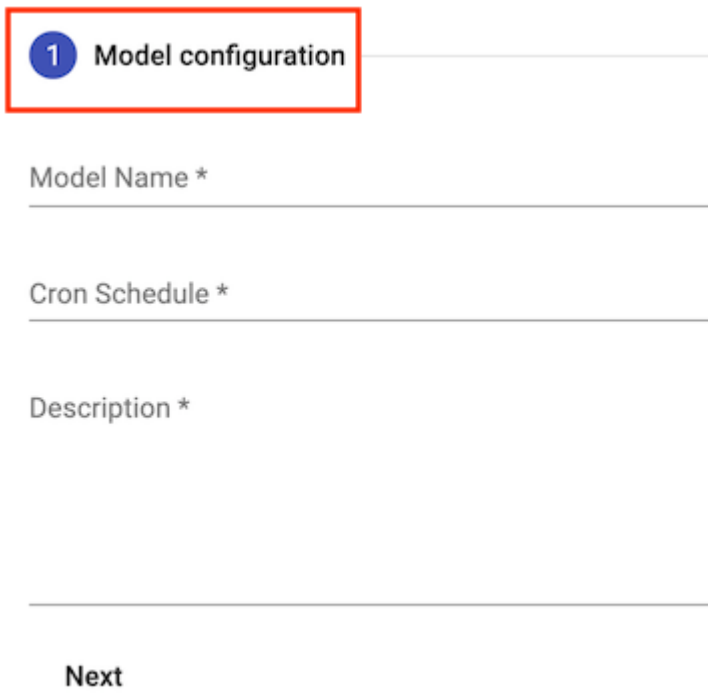
Esta pantalla cuenta con un proceso guiado a través de tres pasos principales, los cuales son: configuración del modelo, selección del modo de diseño y creación de modelo. Estos pasos son detallados a continuación:

- **Configuración del modelo:**

  Permite especificar información del modelo de datos. Estos datos, son los mismos que se muestran en la tabla “Models” explicada anteriormente.

  Imagen 8

  Imagen de configuración del modelo.



  Fuente: Elaboración propia.

  Los datos que se deben ingresar son:

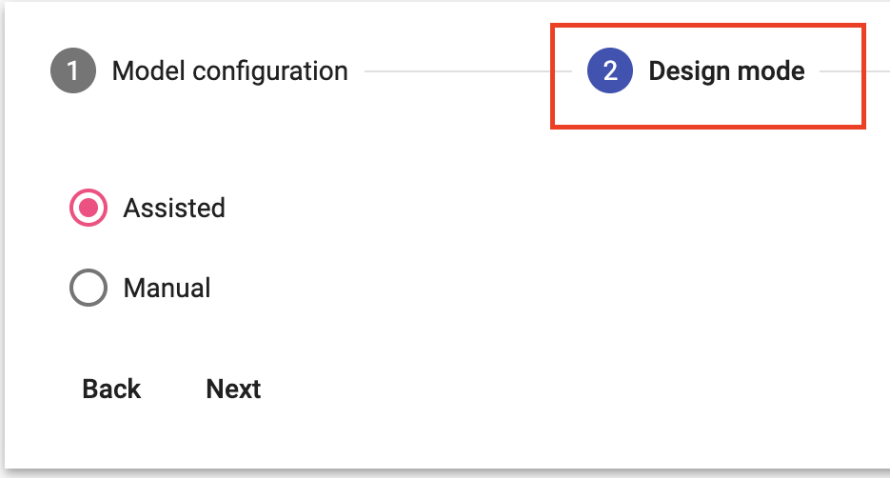
  - **Model Name:** es el nombre del modelo.
  - **Cron Schedule:** es la expresión CRON que indica la frecuencia en la que se debe ejecutar el modelo de datos.
  - **Description:** es la descripción del modelo.

- **Modo de diseño:**

  Permite seleccionar el modo de diseño que se quiere utilizar en el siguiente paso de la creación del modelo.

Imagen 9

Imagen de selección del modo de diseño.

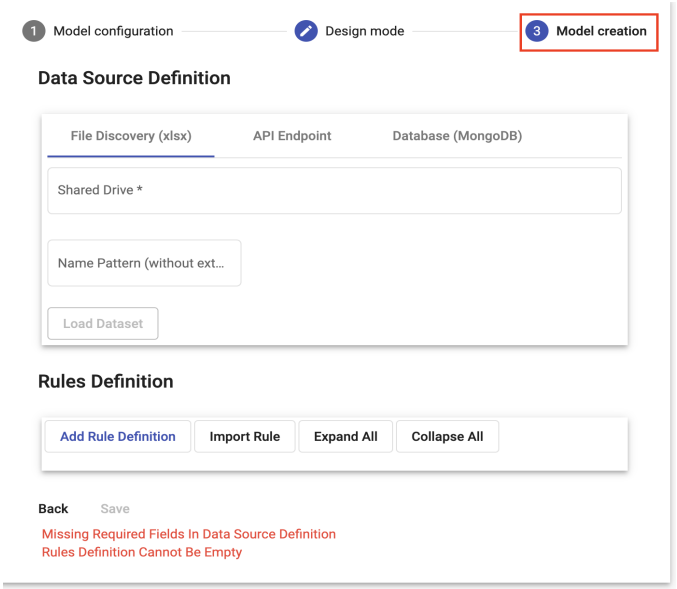


Fuente: Elaboración propia.

- ## Creación del modelo (asistida):

Imagen 10

Imagen de creación del modelo usando modo asistido.



Fuente: Elaboración propia.

Este modo de diseño, cuenta con 2 secciones principales que deben ser completadas.

- **Definición de la fuente de datos:**

  En esta sección, se debe especificar la fuente de datos que se va a utilizar. Inicialmente, el ambiente cuenta con 3 opciones de fuente de datos, sin embargo, estas fuentes son agregadas al ambiente de forma modular. Esto quiere decir, que nuevas fuentes de datos pueden ser habilitadas en el futuro, a través de su respectivo desarrollo. Este último punto, se extiende en la sección "trabajos futuros" del capítulo V de este documento.

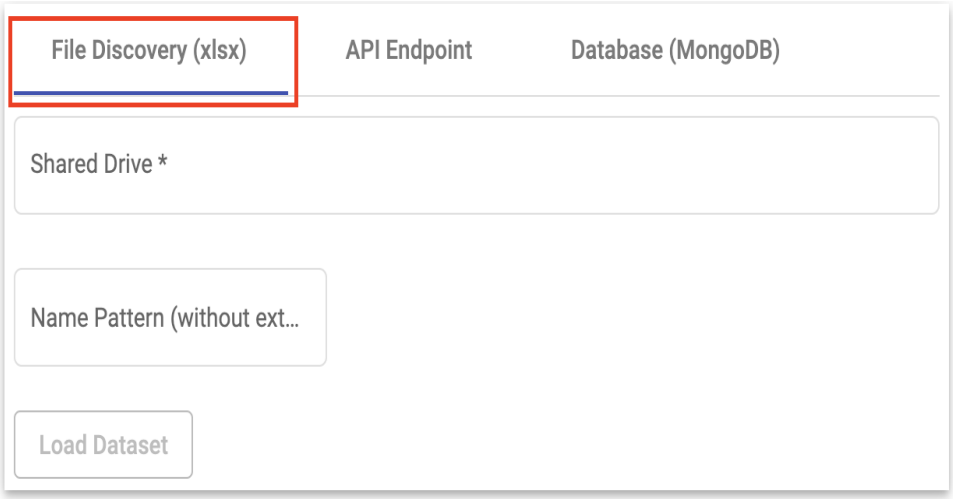
  Las 3 opciones disponibles son las siguientes:

  1. **Descubrimiento de archivos (xlsx):**

     Esta opción permite, inicialmente, identificar archivos nuevos o actualizados con extensión ".xlsx" para luego extraer información de los mismos.

     Imagen 11

     Imagen de fuente datos xlsx.



     Fuente: Elaboración propia.

Los campos disponibles para esta opción de fuente de datos son los siguientes:

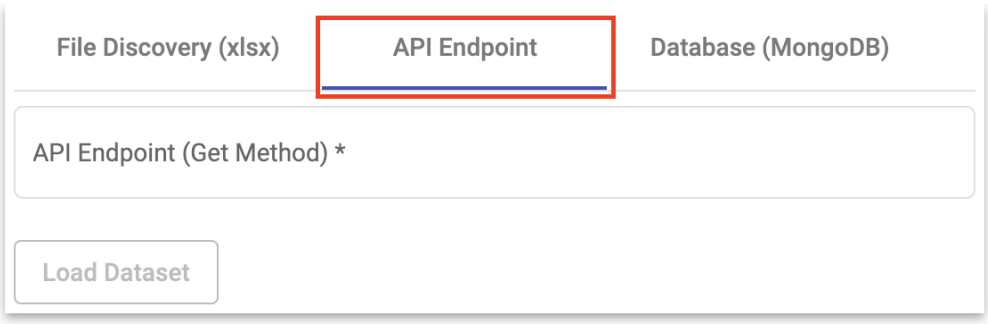
a. **Directorio compartido:** es el directorio que debe ser consultado para descubrir nuevos archivos.
b. **El patrón de nombre:** es el indicador o prefijo de cuáles son los archivos que se deben tomar en cuenta durante el descubrimiento de nuevos archivos ya que en el directorio podrían existir otros documentos que no se deben considerar para este modelo.

2. **API endpoint:**

Esta opción permite obtener los resultados a través de solicitudes a un API.

Imagen 12

Imagen de fuente de datos API endpoint.



Fuente: Elaboración propia.

El campo disponible para esta opción de fuente de datos es el siguiente:

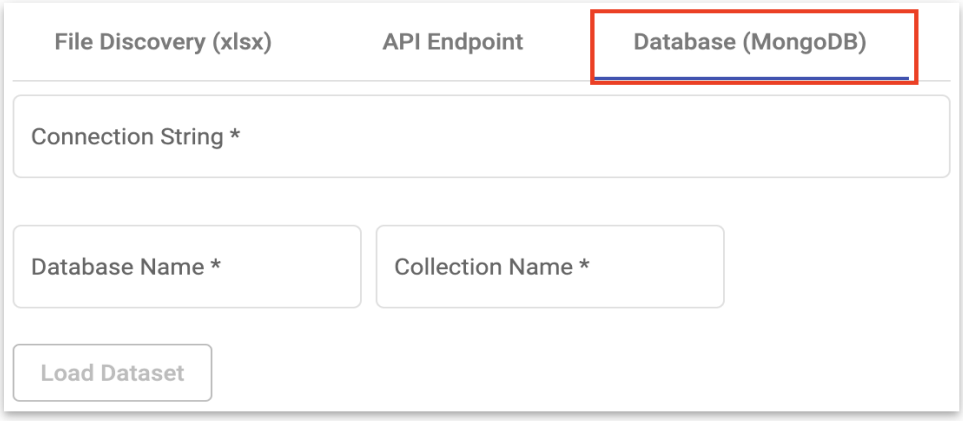
a. API Endpoint: es el endpoint del API donde se debe realizar la consulta, usando el método HTTP GET, para obtener los resultados.

3. **Base de datos (MongoDB):**

Esta opción permite obtener los resultados a través de consultas a una base de datos de MongoDB.

Imagen 13

Imagen de fuente de datos MongoDB



Fuente: Elaboración propia.

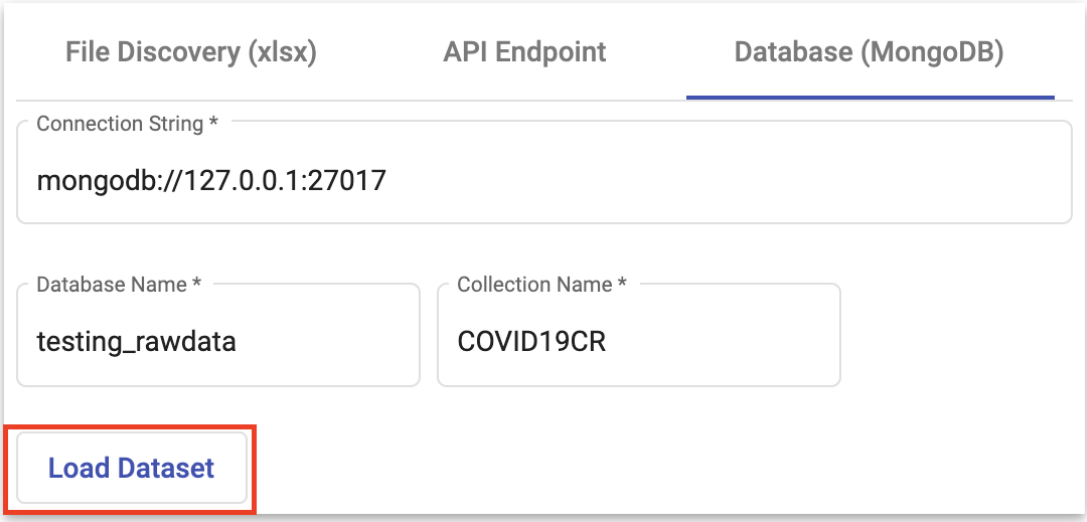
Los campos disponibles para esta opción de fuente de datos son los siguientes:

a. **Cadena de conexión:** es la cadena de conexión a la base de datos que se va a utilizar.
b. **Nombre de la base de datos:** es el nombre de la base de datos que se va a utilizar.
c. **Nombre de la colección:** es el nombre de la colección que se va a utilizar dentro de la base de datos especificada.

Cada fuente de datos, cuenta con un botón llamado “Load Dataset”, el cual tiene como funcionalidad cargar el conjunto de datos. Esto quiere decir, que este botón permite tener una vista previa de los datos y sirve para apoyar la generación de las reglas del modelo. En la siguiente captura, se muestra una fuente de datos seguido de una vista previa de la misma:

Imagen 14

Imagen de botón para cargar el conjunto de datos.

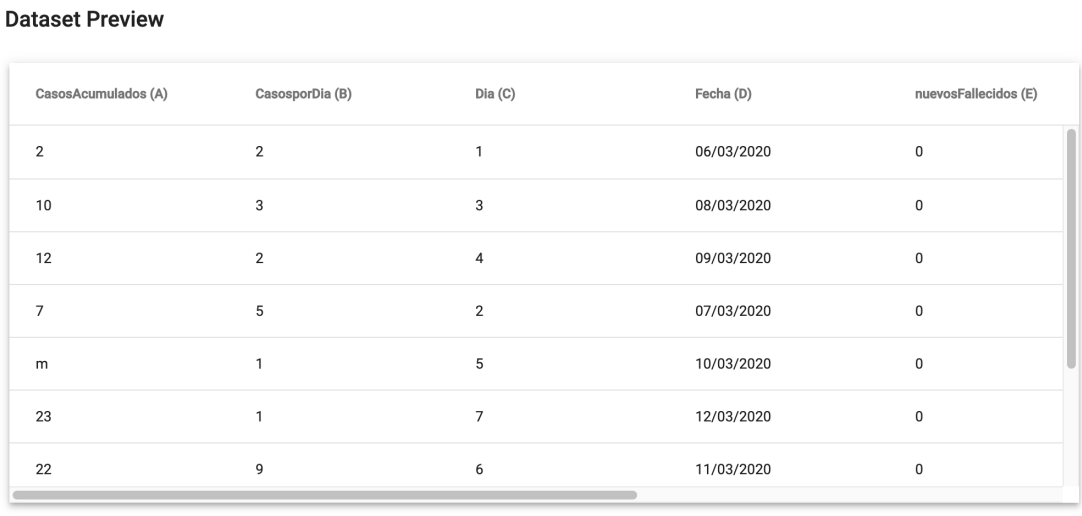


Fuente: Elaboración propia.

Vista previa después de hacer clic al botón “Load Dataset”:

Imagen 15

Imagen de vista previa de los datos presentes en el conjunto de datos.



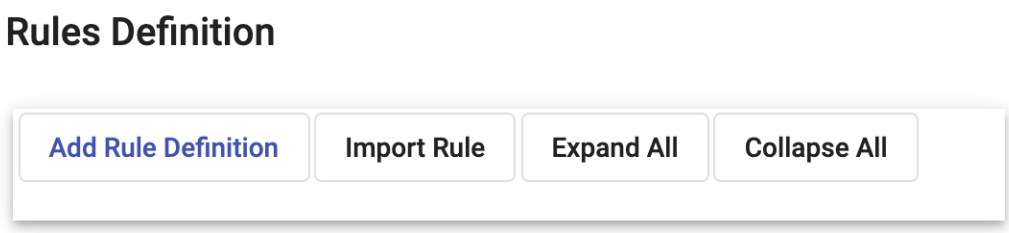
Fuente: Elaboración propia.

- **Definición de las reglas:**

  En esta sección, se crean las reglas que van a ser agregadas al modelo. Una regla es una acción que se debe aplicar a los datos durante la ejecución del modelo.

  Imagen 16

  Imagen de definición de las reglas.



  Fuente: Elaboración propia.

  La sección de definición de las reglas, cuenta con 4 botones que se detallan a continuación:

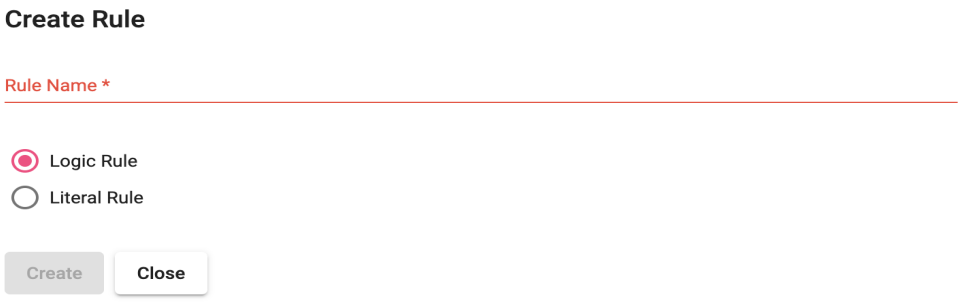
  1. **Agregar definición de la regla:**

     Permite agregar una nueva regla asistida. Una regla asistida, es un conjunto de instrucciones básicas que deben ser aplicadas a un modelo.

     Al hacer clic en este botón, se muestra una ventana emergente en la que se debe establecer un nombre para la regla y seleccionar el tipo de regla que desea crear.

     Imagen 17

     Imagen de crear regla.



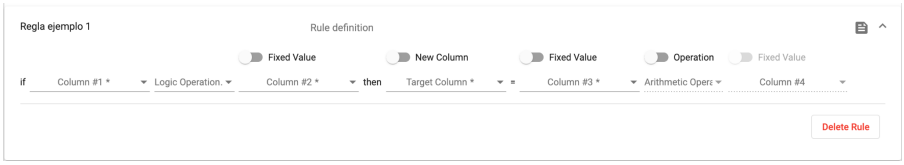
     Fuente: Elaboración propia.

Los tipos de regla disponibles son:

a. **Regla lógica:** permite crear operaciones lógicas entre distintas columnas del conjunto de datos.

Imagen 18

Imagen de regla lógica.

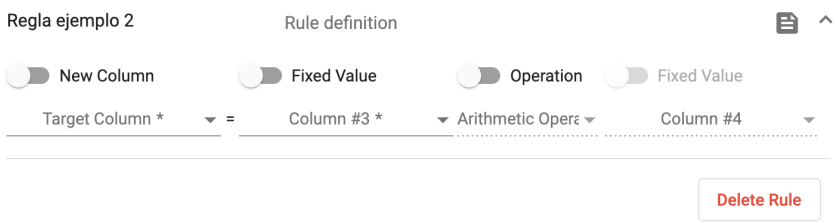


Fuente: Elaboración propia.

b. **Regla literal:** permite crear operaciones aritméticas entre distintas columnas del conjunto de datos.

Imagen 19

Imagen de regla literal.



Fuente: Elaboración propia.

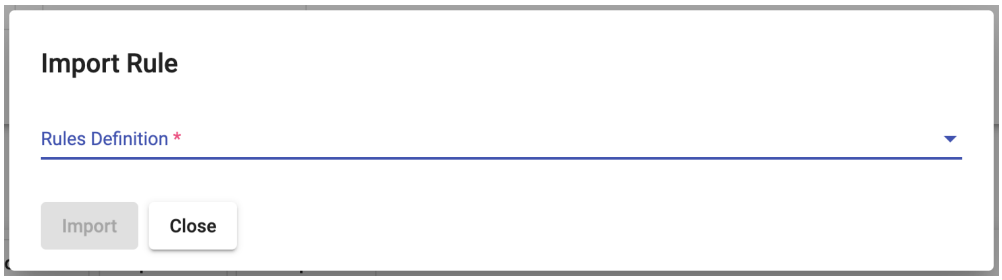
Cada tipo de regla, cuenta con un botón llamado "Delete Rule", el cual tiene como objetivo borrar una regla cuando ya no es necesaria en el modelo.

2. **Importar regla:**

Permite importar reglas existentes en el ambiente con el fin de reutilizarlas. Al hacer clic en este botón, se muestra una ventana emergente en la que se debe seleccionar la regla que se desea importar.

Imagen 20

Imagen de importar regla.



Fuente: Elaboración propia.

La selección se hace utilizando una lista desplegable que muestra las reglas compatibles que pueden ser importadas a este modelo.

Imagen 21

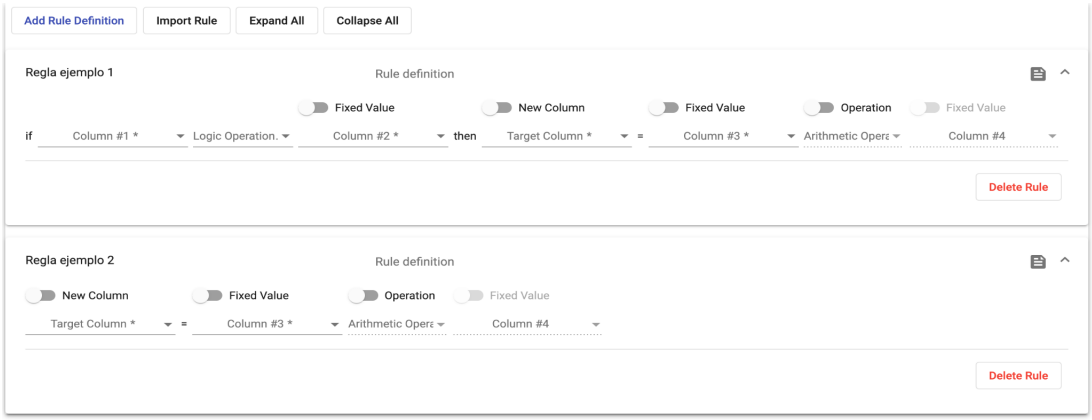
Imagen de selección de la regla a importar.

Fuente: Elaboración propia.

3. **Expandir todo:** permite expandir todas las reglas creadas.

Imagen 22

Imagen de opción de expandir todo.



Fuente: Elaboración propia.

4. **Colapsar todo:** permite colapsar todas las reglas creadas.

Imagen 23

Imagen de opción de colapsar todo.

Fuente: Elaboración propia.

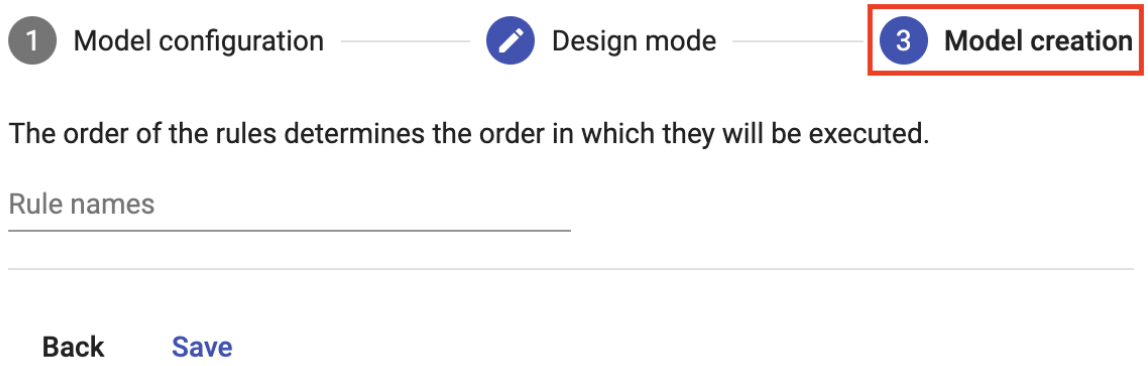
En el anexo 4, se puede observar un modelo completo creado con el modo de diseño asistido. Se recomienda revisar este anexo ya que sirve como un excelente complemento para comprender la propuesta.

- **Creación del modelo (manual):**

El modo de diseño manual, está orientado a usuarios con un conocimiento técnico más amplio ya que permite crear reglas con instrucciones en código, utilizando el lenguaje de programación Python. Los lenguajes soportados pueden ser expandidos en el futuro. Esto último, se extiende en la sección “trabajos futuros” del capítulo V de este documento.

Imagen 24

Imagen de creación del modelo usando modo manual.



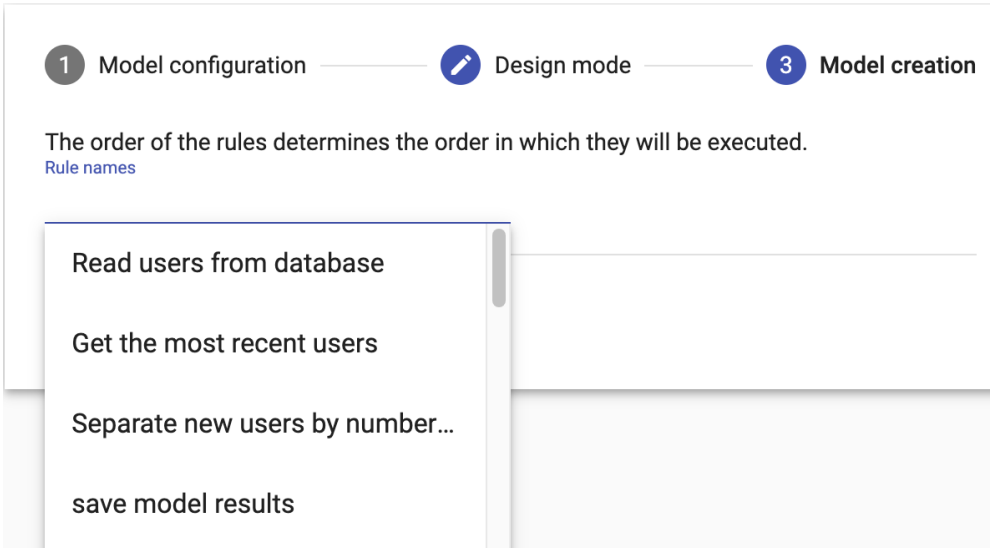
Fuente: Elaboración propia.

Esta pantalla permite seleccionar reglas, que han sido creadas previamente, para agregarlas al modelo. El modo de creación de estas reglas será detallado más adelante. Como se puede deducir, las reglas del modo manual son diferentes a las mencionadas en el modo de diseño asistido y su flexibilidad es mucho más amplia. Con estas reglas, el usuario tiene total control de las instrucciones que el modelo debe ejecutar. Es valioso mencionar, que en la sección "trabajos futuros" se menciona un aspecto importante, en términos de seguridad e infraestructura, que deben ser considerados al utilizar este modo de diseño.

La selección de las reglas que se deben agregar al modelo, se hace a través de una lista desplegable. Es importante mencionar, que el orden en el que las reglas son agregadas también es el orden en el que van a ser ejecutadas.

Imagen 25

Imagen de selección de reglas para modo manual.

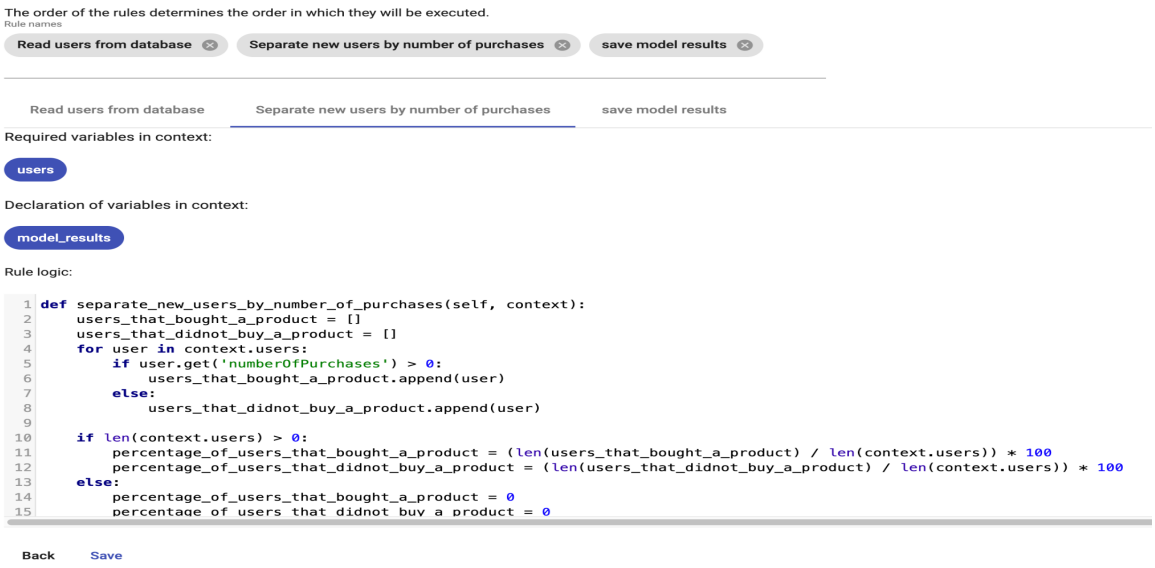


Fuente: Elaboración propia.

Una vez que las reglas han sido seleccionadas, se muestran los detalles de las mismas con el objetivo de ayudar al usuario a crear el modelo correctamente, de acuerdo a los requerimientos existentes.

Imagen 26

Imagen de la vista de detalles de las reglas seleccionadas.



Fuente: Elaboración propia.

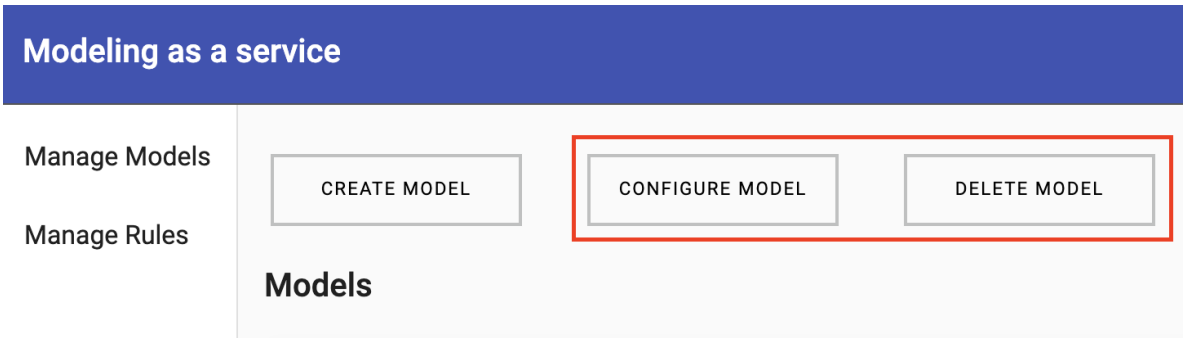
La información mostrada para cada regla es:

- **Variables requeridas en contexto:** se refiere a variables que el código de la regla necesita antes de ejecutarse. Estas variables deben ser declaradas por alguna otra regla previamente.
- **Declaración de variables en contexto:** se refiere a las variables que son declaradas por la regla y que pueden ser utilizadas por otra regla posteriormente.
- **Lógica de la regla:** muestra el código o instrucciones que debe ejecutar cada regla para cumplir con su propósito.

Seguidamente, se menciona el propósito de los botones para configurar y borrar modelos.

Imagen 27

Imagen de botones configurar y borrar que permiten gestionar modelos.



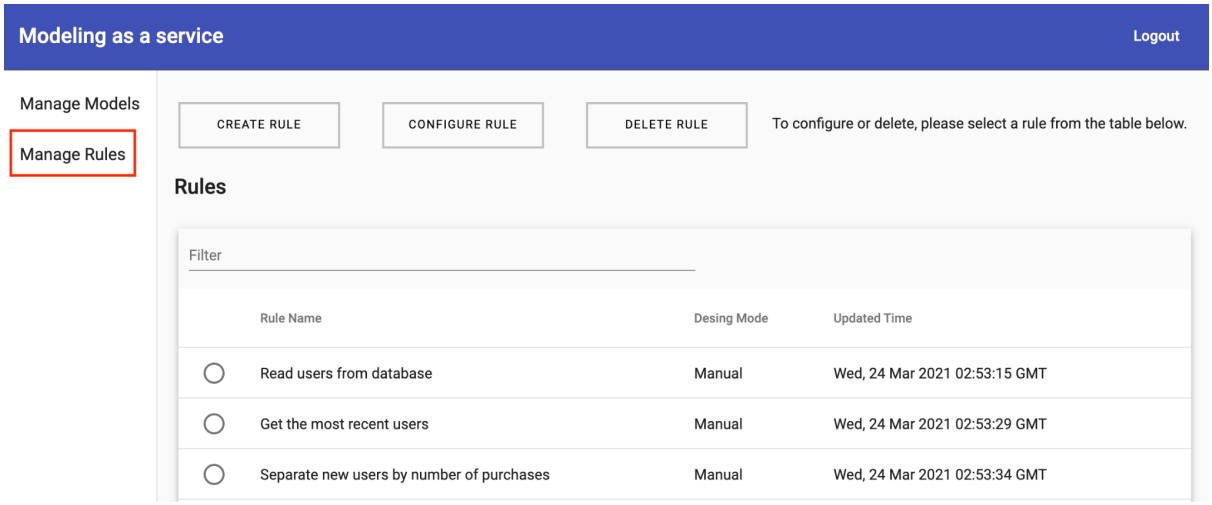
Fuente: Elaboración propia.

- **Configurar:** permite editar el modelo seleccionado, utilizando las mismas pantallas para crear un modelo nuevo.
- **Borrar:** permite borrar el modelo seleccionado.

Adicionalmente, se cuenta con una pantalla para gestionar las reglas del modo manual. Esta pantalla muestra información relevante de las reglas existentes, así como también botones para ejecutar distintas acciones. Al igual que la tabla de modelos, la primera columna de la tabla llamada “Rules” es un botón de radio que permite seleccionar la regla que se quiere gestionar. Dicha pantalla, puede ser observada en la siguiente captura:

Imagen 28

Imagen de vista para gestionar reglas.



Fuente: Elaboración propia.

Las columnas mostradas en la tabla son las siguientes:

- **Nombre de la regla:** muestra el nombre de la regla asignado por el usuario.
- **Modo de diseño:** muestra el modo de diseño utilizado para crear la regla.
- **Updated Time:** Indica la última vez que la regla fue actualizada.

El botón “create rule”, permite crear una nueva regla manual que puede ser luego agregada a un modelo, utilizando el modo de diseño manual.

Imagen 29

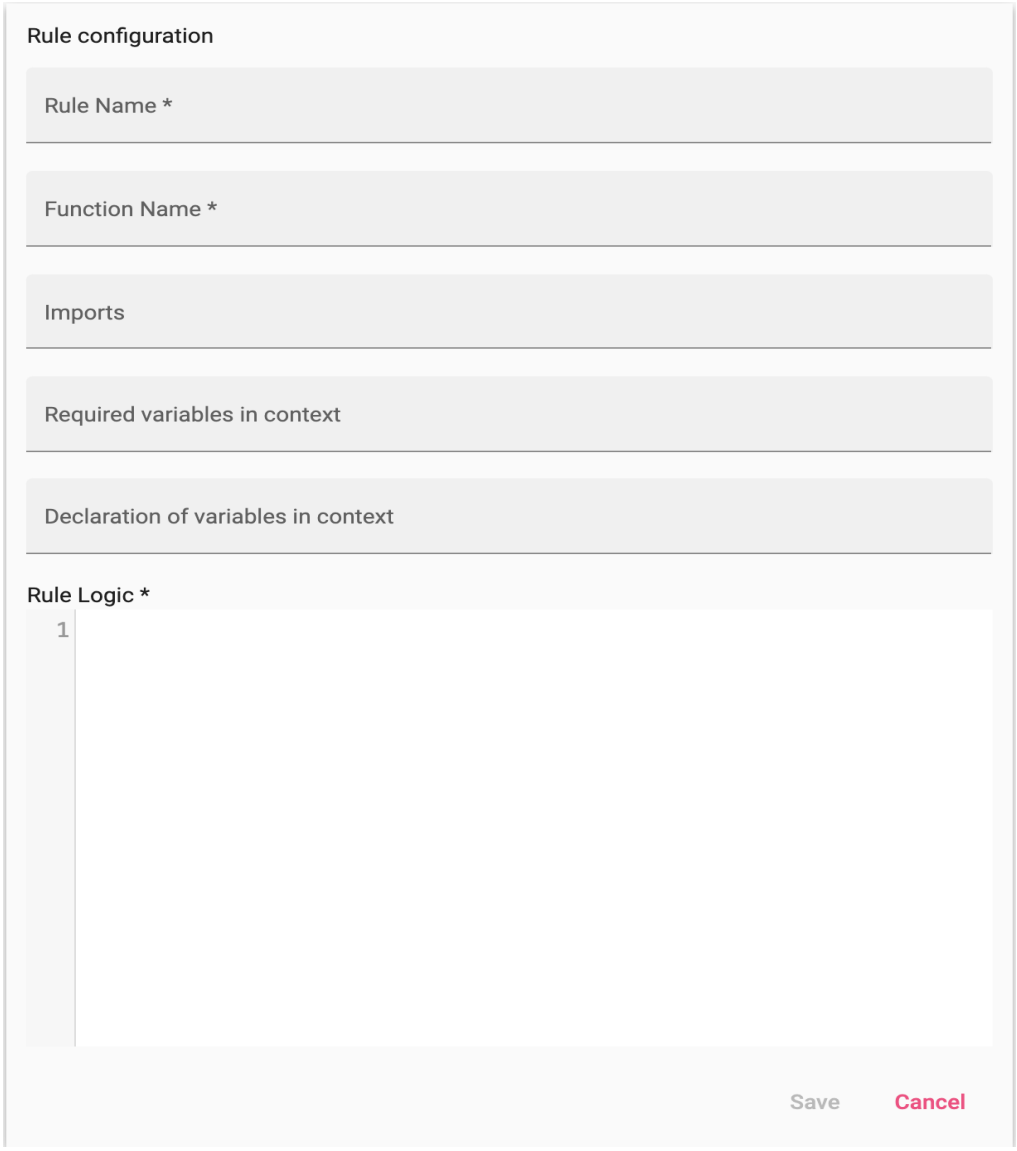
Imagen de botón para crear regla.

Fuente: Elaboración propia.

Seguidamente, se puede observar la pantalla mostrada al utilizar el botón “create rule”:

Imagen 30

Imagen de vista para configurar regla.



Fuente: Elaboración propia.

El propósito de los campos, de esta pantalla, es el siguiente:

- **Nombre de la regla:** es el nombre de la regla.
- **Nombre de la función:** es un nombre representativo para la acción que la regla va a desempeñar.

- **Importaciones:** son las dependencias o requerimientos que deben ser resueltos antes de ejecutar la regla.
- **Variables requeridas en contexto:** permite ingresar las variables que el código de la regla necesita antes de ejecutarse.
- **Declaración de variables en contexto:** permite ingresar las variables que son declaradas por la regla.
- **Lógica de la regla:** permite ingresar el código o instrucciones que debe ejecutar cada regla para cumplir con su propósito.

A continuación, se menciona el propósito de los botones para configurar y borrar reglas.

Imagen 31

Imagen de botones configurar y borrar que permiten gestionar reglas.

Fuente: Elaboración propia.

- **Configurar:** permite editar la regla seleccionada, utilizando la misma pantalla para crear una regla nueva.
- **Borrar:** permite borrar la regla seleccionada.

Como herramienta de extracción de datos del ambiente, se propone utilizar Dremio. Esta herramienta, le brinda al ambiente la posibilidad de incorporar nuevas tecnologías de almacenamiento de datos en el futuro, sin afectar el modo de consumo de los datos. Esto es beneficioso para el ambiente de modelado de datos ya que le permite mantener su enfoque como servicio, donde el usuario no debe preocuparse por el funcionamiento interno del sistema.

Como herramienta de visualización de datos, se propone utilizar Power BI Desktop. Esta herramienta, es muy beneficiosa para gestionar y visualizar datos provenientes de los modelos creados en el ambiente de modelado de datos. Así mismo, Power BI Desktop se puede conectar a Dremio para extraer la información necesaria, y así, realizar distintos análisis predictivos y descriptivos.

3. Validación de la propuesta

El término validar se refiere a "*Dar fuerza o firmeza a algo, hacerlo válido.*" (Real Academia Española, s.f., definición 1).

Con el objetivo de validar la propuesta, se realizan pruebas virtuales locales que permiten conocer el estado, tiempo y costo del proceso de modelado de datos. Estas pruebas se separan en dos escenarios, uno utilizando el ambiente de modelado de datos y otro creando los modelos directamente en Power BI Desktop sin hacer uso del ambiente. Las pruebas consisten en la creación de un modelo simple que permite obtener información extra, a partir de datos existentes, utilizando columnas calculadas.

Tabla 4

Tabla comparativa del tiempo requerido para crear un modelo con datos listos para ser utilizados.

Escenario	Conjunto de datos pequeño (15 columnas y 448 filas)	Conjunto de datos mediano (200 columnas y 448 filas)
Utilizando el ambiente	3 minutos aproximadamente	3 minutos aproximadamente
Sin utilizar el ambiente	2.5 minutos aproximadamente	5 minutos aproximadamente

Fuente: Elaboración propia.

Como se observa en la tabla anterior, el tiempo requerido para crear un modelo es diferente dependiendo si se usa o no el ambiente de modelado de datos. Los escenarios de estas pruebas se detallan a continuación:

- **Utilizando el ambiente:**

  Al crear un nuevo modelo, el ambiente escanea la fuente de datos con el objetivo de identificar las columnas válidas que pueden ser utilizadas. Al solo identificar las columnas válidas, sin realizar ningún tipo de conversión de datos, la creación del modelo no se ve prácticamente afectado por el tamaño del conjunto de datos.

  Una vez que el modelo es creado, se ejecuta el mismo para realizar todo el procesamiento y aplicar las reglas definidas por el usuario. Para ser justos con las pruebas, al tiempo aproximado de esta prueba se le sumó el tiempo requerido para ejecutar el modelo y tener los datos listos para ser utilizados.

- **Sin utilizar el ambiente:**

  Al no utilizar el ambiente y solo usar Power BI Desktop, todo el procesamiento se da en la máquina local donde se está ejecutando la aplicación. Esto produce que la máquina local ejecute todos los cálculos del modelo de acuerdo a las acciones que el usuario realiza.

  Al crear columnas calculadas, estas se completan con los datos correspondientes después de ser definidas por el usuario. Con esto último, se experimentó una degradación del rendimiento al utilizar un conjunto de datos mediano y produjo retrasos para completar el modelo requerido por la prueba.

Dado esto, se identifica que el rendimiento es muy similar cuando el conjunto de datos es pequeño, siendo Power Bi Desktop ligeramente más rápido ya que el usuario no debe esperar a que se ejecute el modelo basado en una calendarización, como es el caso del ambiente de modelado de datos. Sin embargo, el rendimiento de Power Bi Desktop se ve afectado cuando el conjunto de datos va creciendo debido a que los cálculos deseados en el modelo se procesan en la máquina local, conforme el usuario interactúa con los mismos. Es

importante mencionar, que al igual que cualquier software, el rendimiento de la herramienta Power Bi Desktop también se asocia a la capacidad de procesamiento del hardware utilizado.

El ambiente de modelado, requiere un menor tiempo para crear el modelo cuando se usan conjuntos de datos más grandes. Además, el ambiente muestra un menor tiempo requerido para realizar el procesamiento de los cálculos o reglas definidos en el modelo, lo cual permite tener los datos listos para ser utilizados en menor tiempo.

Para estas pruebas, se utilizó una computadora con las siguientes características:

- CPU: Intel i7-6700HQ.
- Memoria: 16 GB DDR4.
- Almacenamiento: Unidad de estado sólido (SSD).

Con la validación anterior, se identifica que existe una mejora en tiempo al usar el ambiente de modelado de datos. Esta mejora es notable al compararse con el proceso o alternativa previamente existente para crear modelos.

Para validar la factibilidad de la ejecución del ambiente de modelado de datos en la nube, se procede a instalar el ambiente en Azure. La instalación se realiza utilizando contenedores Linux y servicios de MongoDB Atlas. Después de la validación, no se encontró ningún inconveniente al instalar el ambiente en esta nube y se logra confirmar su factibilidad. Basado en esto, ahora se puede contar con un ambiente de modelado de datos como servicio que puede ser ejecutado completamente en la nube, lo cual beneficia el estado actual del proceso de modelado de datos a través de una nueva alternativa.

Tomando como base la estimación de costos del anexo 1, se valida el costo aproximado de la implementación del ambiente de modelado de datos en una empresa mediana de aproximadamente 100 empleados. Como se aprecia en el anexo, el ambiente representa una solución más económica ya que no se debe asumir el costo de otras soluciones y licencias.

Además, Power Bi Desktop es una herramienta gratuita que puede ser utilizada como parte del ambiente de modelado de datos sin problemas, de acuerdo a la licencia actual. Después de esta validación, se puede identificar que existe una mejora con respecto a costos en el proceso de modelado de datos al utilizar el ambiente.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

1. Conclusiones

1. Los objetivos del proyecto se cumplieron satisfactoriamente a través de instrumentos como consultoría, revisión de literatura y encuestas.
2. No se logró identificar ningún software que permita crear modelos de datos como servicio. Por lo tanto, ahora ya se cuenta con un nuevo ambiente o solución de modelado de datos como servicio listo para ser implementado por las organizaciones.
3. El ambiente de modelado de datos es amigable con cualquier fuente de datos, sin embargo, cada nueva fuente debe ser habilitada a través del desarrollo de un nuevo módulo en el sistema.
4. Habilitó la oportunidad de poner en práctica conocimientos existentes que ahora son más sólidos gracias a la experiencia obtenida en este proyecto.
5. Al adoptar el ambiente de modelado de datos, el costo de infraestructura y mantenimiento que debe asumir la organización es bajo comparado a otras alternativas. Asimismo, el costo de infraestructura asociado varía dependiendo de las necesidades de la organización. Además, los costos de la infraestructura se ven directamente impactados por la capacidad de procesamiento y almacenamiento requerida por la organización.
6. El ambiente desarrollado de modelado de datos cuenta con las tecnologías necesarias para operar correctamente ya sea con pocas o grandes cantidades de datos.

2. Limitaciones

Entre las limitaciones del proyecto, se encuentra que actualmente no se ha logrado identificar ningún software que permita crear modelos de datos como servicio. Al no haberse logrado identificar ninguna solución o software similar, se dificulta la obtención de información sobre soluciones similares que pueden ser estudiadas para comprender cómo el problema ha sido abordado desde otra perspectiva. Para resolver esta limitación, se procedió a hacer uso de consultoría con expertos en el tema de base de datos y diseño de software que permitió generar una mejor solución para el ambiente de modelado de datos.

También, al no lograr identificar soluciones similares que permitan crear modelos de datos como servicio, se dificulta la realización de pruebas para comparar el uso del ambiente de modelado de datos contra soluciones ya existentes. Para resolver esta limitación, se hacen pruebas utilizando herramientas existentes de inteligencia de negocios que son utilizadas para obtener información valiosa a raíz de los datos de las organizaciones.

Sin embargo, resulta difícil ejecutar pruebas precisas entre el ambiente de modelado de datos y otras herramientas ya que su modo de funcionamiento es distinto. Esto ocurre, debido a que el ambiente de modelado de datos permite generar modelos como servicio que se actualizan automáticamente del lado del servidor, con el fin de mantener los datos actualizados y listos para su consumo. Por otra parte, muchas herramientas de inteligencia de negocios funcionan a través de consultas directas a fuentes de datos, generando un reporte o gráfico con la información filtrada.

Al ser los modos de funcionamiento distintos, solo es posible comparar precisamente el reporte o salida final del modelo de datos. Sin embargo, esta comparación solo se realiza en términos de exactitud de la información recuperada y el tiempo de respuesta que se necesita para completar la solicitud y poder observar los datos.

Adicionalmente, las preguntas de negocio en las organizaciones pueden llegar a ser muy específicas y, a su vez, requerir un conjunto de datos con mucha información que

permita llegar a una conclusión o respuesta certera a través de un modelo de datos. De modo que se vuelve una limitante para el proyecto ya que se necesita un conjunto de datos muy elaborado que requiere muchos recursos, lo cual se sale de las posibilidades del autor y de los objetivos del proyecto. Como alternativa, se decide utilizar datos del INEC y del Observatorio del desarrollo de la Universidad de Costa Rica que pueden ser aprovechados para obtener información valiosa y ejecutar las pruebas necesarias.

3. Trabajos futuros

Inicialmente, se recomienda habilitar un inicio de sesión único o SSO, por sus siglas en inglés. Esto permite que la aplicación web, del ambiente de modelado de datos, identifique a cualquier usuario que utiliza el sistema utilizando una misma identificación en toda la organización. Sin embargo, si no se desea utilizar un SSO, se recomienda habilitar un módulo para registrar nuevos usuarios en el sistema.

Actualmente, el ambiente de modelado de datos cuenta con soporte para distintas fuentes de datos. No obstante, nuevas fuentes de datos pueden ser habilitadas en cualquier momento. Se recomienda habilitar nuevas fuentes de datos, de acuerdo a las necesidades de la organización.

También, para la creación de las reglas manuales, se utiliza el lenguaje de programación Python. Aún así, es posible agregar soporte para nuevos lenguajes de programación, si así lo requiere la organización.

Adicionalmente, se recomienda habilitar un mecanismo o módulo que permita crear y gestionar reglas privadas. De este modo, se puede habilitar reglas de acciones críticas como lo es guardar los resultados de la ejecución de un modelo en la base de datos, sin comprometer aspectos importantes del ambiente de modelado de datos.

Además, es altamente recomendable ejecutar los modelos en contenedores. De esta forma, se evita comprometer el servidor donde se está ejecutando el código Python ingresado por el usuario, al momento de crear reglas manuales.

REFERENCIAS

REFERENCIAS

Arias, Fidias (2006). El proyecto de investigación: Introducción a la metodología científica. (5º. ed.) Caracas - Venezuela: Episteme

Bernal, C. (2010). Metodología de la investigación para Administración. Economía, Humanidades y Ciencias Sociales. (3nda ed.). México, D.F: Pearson

Calzada, L., & Abreu, J. L. (2009, septiembre). El impacto de las herramientas de inteligencia de negocios en la toma de decisiones de los ejecutivos. Instituto de estudios superiores - spentamexico. http://www.spentamexico.org/v4-n2/4(2)%2016-52.pdf

Campos, E. M. (2004). INTERNET Y SOCIEDAD: RELACIÓN Y COMPROMISO DE BENEFICIOS COLECTIVOS E INDIVIDUALES. Revista Digital Universitaria, 10.

Castro, J. (2015, Agosto 12). que-es-la-inteligencia-de-negocios. Obtenido de corponet: https://blog.corponet.com.mx/que-es-la-inteligencia-de-negocios

Castro Rozo, F. E. (2013, noviembre). Indicadores de gestión para la toma de decisiones basada en Inteligencia de Negocios. Sistema de Revistas Científicas - Universidad Distrital Francisco José De Caldas. https://revistas.udistrital.edu.co/index.php/tia/article/view/4639/7094

Cordero, C. C. (4 de noviembre de 2014). Usuarios generan 2,5 quintillones de bytes en datos diarios, dice comScore. Obtenido de https://www.elfinancierocr.com/tecnologia/usuarios-generan-25-quintillones-de-bytes-en-datos-diarios-dice-comscore/YMCBM6XIBJC7FG7ZQXUYZIL7CY/story/

Davenport, T. H. (Enero de 2013). https://www.sas.com/es_mx/insights/big-data/what-is-big-data.html. Obtenido de www.sas.com: https://www.sas.com/es_mx/whitepapers/bigdata-bigcompanies-106461.html

Delfín Pozos, F. L., & Acosta Márquez, M. P. (2016). *Importancia y análisis del desarrollo empresarial.* Veracruz (México): Instituto de la Contaduría Pública – Universidad Veracruzana, Xalapa, Veracruz (México).

González, G. (2020, 6 abril). Investigación diagnóstica: características, técnicas, tipos, ejemplos. Lifeder. https://www.lifeder.com/investigacion-diagnostica/

Jacobson, Ralph. (24 April de 2013). *IBM Consumer Products Industry Blog*. Obtenido dehttps://www.copibec.ca/medias/files/PDF%20(EN)/copibec-copyright-information-flow.pdf

Landeau, R. (2007). Elaboración de Trabajos de Investigación. Caracas: Editorial Alfa.

Lewinter, H. (2016). Why Data Is Important To Your Business Success. Obtenido de https://www.nimble.com/blog/why-data-is-important-to-your-business-success/

Meyer, C. (2015, 22 mayo). Big Data en contexto. ELMUNDO. https://www.elmundo.es/economia/2015/05/22/555f00c122601db75d8b4584.html

Ocegueda, C. (2004). Metodología de la investigación. Métodos, técnicas y estructuración de trabajos académicos (Vol. 2).

Olivera, X. (16 de Noviembre de 2016). La importancia de los datos en tiempo real para cuidar el gasto empresarial. Obtenido de http://spendmatters.com/mx-latam/about/:

http://spendmatters.com/mx-latam/la-importancia-de-los-datos-en-tiempo-real-para-cuidar-el-gasto-empresarial/

Parms, J. (2015). Datafloq. Obtenido de 6 Benefits of Data Modeling in the Age of Big Data: https://datafloq.com/read/6-benefits-data-modeling-in-the-age-of-big-data/1479

Project Management Institute. (2008) . GUÍA DE LOS FUNDAMENTOS PARA LA DIRECCIÓN DE PROYECTOS, 4ª. ed. Estados Unidos: PMI Publications

Rouse, M. (Enero de 2015). techtarget. Obtenido de techtarget: https://searchdatacenter.techtarget.com/es/definicion/Base-de-datos-relacional

REAL ACADEMIA ESPAÑOLA: Diccionario de la lengua española, 23.ª ed., [versión 23.4 en línea]. https://dle.rae.es [4/8/2021].

Zapata, D. (2006). Restauración de la Plaza "Delia Ávila de Zapata", de la Urbanización El Valle, Municipio Bermúdez Estado Sucre. Trabajo de Grado. Universidad de Oriente.

GLOSARIO

Backend: Es la parte trasera del sistema de información que se encarga del procesamiento lógico. Ayuda a que la interfaz de usuario funcione correctamente.

UI: Es la interfaz de usuario por sus siglas en inglés (user interface). La interfaz es el diseño que utilizan los usuarios para interactuar con el sistema.

Big data: Se refiere a grandes cantidades de datos tanto estructurados como no estructurados que tienen un grado importante de complejidad y crecen constantemente.

Data warehouse: Es un almacén electrónico donde se integra y centraliza datos de diferentes fuentes con el fin de ser procesados luego, generalmente para apoyar análisis empresariales.

Software: Es un programa informático o conjunto de rutinas que permiten a una computadora realizar tareas.

Hardware: Es un elemento físico que forma parte de una computadora.

JDBC driver: Es un software que permite, a una aplicación desarrollada utilizando el lenguaje de programación Java, interactuar con una base de datos.

Cluster: Es el conjunto de múltiples computadores que trabajan para completar un objetivo común.

SQL: Es un lenguaje de consulta estructurado por sus siglas en inglés (structured query language). Es un lenguaje destinado para gestionar bases de datos.

JSON: Es un estándar de formato de archivo que utiliza texto legible para los humanos y que a su vez representan objetos de datos.

Framework: Es un marco de trabajo, un conjunto de prácticas y conceptos que se enfocan en resolver un objetivo en particular.

CPU: Es la unidad central de procesamiento por sus siglas en inglés (central processing unit). Es el componente de una computadora que se encarga del procesamiento.

Insights: Es el descubrimiento de algo que no es obvio o que no es percibido fácilmente.

As a service: Es un término utilizado para describir soluciones a problemas basado en computación en la red, usualmente internet.

Bit: es la unidad mínima de información en informática, la cual contiene valores binarios.

Bytes: Es una medida de computadora equivalente a un conjunto de 8 bits.

Terabyte (TB): Es una medida de memoria de computadora que es equivalente a 1 billón de bytes.

Gigabyte (GB): Es una medida de memoria de computadora que es equivalente a 1 1 millón de bytes.

CRON: Es una expresión que permite establecer un tiempo. Se utiliza para planificar la ejecución de una tarea automatizada.

XLSX: Es una extensión de archivo de versiones de Microsoft® Excel® modernas.

API: Es una serie de procedimientos que pueden ser utilizados por otro software.

Endpoint: Es la dirección de un API que responde a una petición.

HTTP: Es un protocolo que permite intercambio de datos en la web.

GET: Es un método HTTP que permite recuperar información.

MongoDB: Es una base de datos no relacional orientada a documentos.

Angular: Es un framework para desarrollar aplicaciones web de una sola página.

Python: Es un lenguaje de programación interpretado, multiplataforma y sencillo de leer.

Single Sign-On (SSO): Es un mecanismo de autenticación que permite a los usuarios de una organización acceder a varios sistemas utilizando una sola identificación.

Unidad de estado sólido (**SSD):** Es un dispositivo de almacenamiento de datos significativamente más rápido.

Power BI Desktop®: Es una herramienta que permite crear informes y gestionar datos de una manera avanzada.

DDR4: Es un tipo de memoria de computadora.

Intel i7-6700HQ: Es un procesador de corporación Intel.

Linux: Es un sistema operativo de código abierto.

Azure: Es un servicio de computación en la nube.

Dremio: Es una plataforma de gestión de grandes cantidades datos.

ANEXOS

Anexo 1

Estimación de costos			
Rubro	Costo anual estimado	Costo estimado por 3 años	Notas
Costo promedio de un ingeniero para mantenimiento del ambiente	$ 18 709	$ 56 127	
Costo de licencias	$ 0	$ 0	
Máquina virtual en Azure (servidor web y API)	$ 366.24	$ 1 098.72	Incluye 32GB de almacenamiento SSD
Instancias de contenedores (procesamiento de modelos de datos)	$ 776.76	$ 2 330.28	
Almacenamiento en base de datos	$ 3 280.32	$ 9 840.96	2 TB
Total	$ 23 132.32	$ 69 396.96	

Anexo 2

Validación de otras alternativas				
Nombre de la solución	Costo anual estimado	Costo estimado por 3 años	Costo total (incluye el ingeniero y la maquina virtual)	Notas
Costo promedio de un ingeniero para mantenimiento del ambiente	$ 18 709	$ 56 127		
Máquina virtual en Azure (gestión de datos)	$ 366.24	$ 1 098.72		
Amazon Redshift	$ 20 301.72	$ 60 905.16	$ 118 130.88	2TB de inserción y consumo por mes
Google BigQuery	$ 24 474.6	$ 73 423.8	$ 130 649.52	2TB de inserción y consumo por mes
SnowFlake	N/A	N/A	N/A	Requiere un registro de la empresa interesada para acceder a los precios

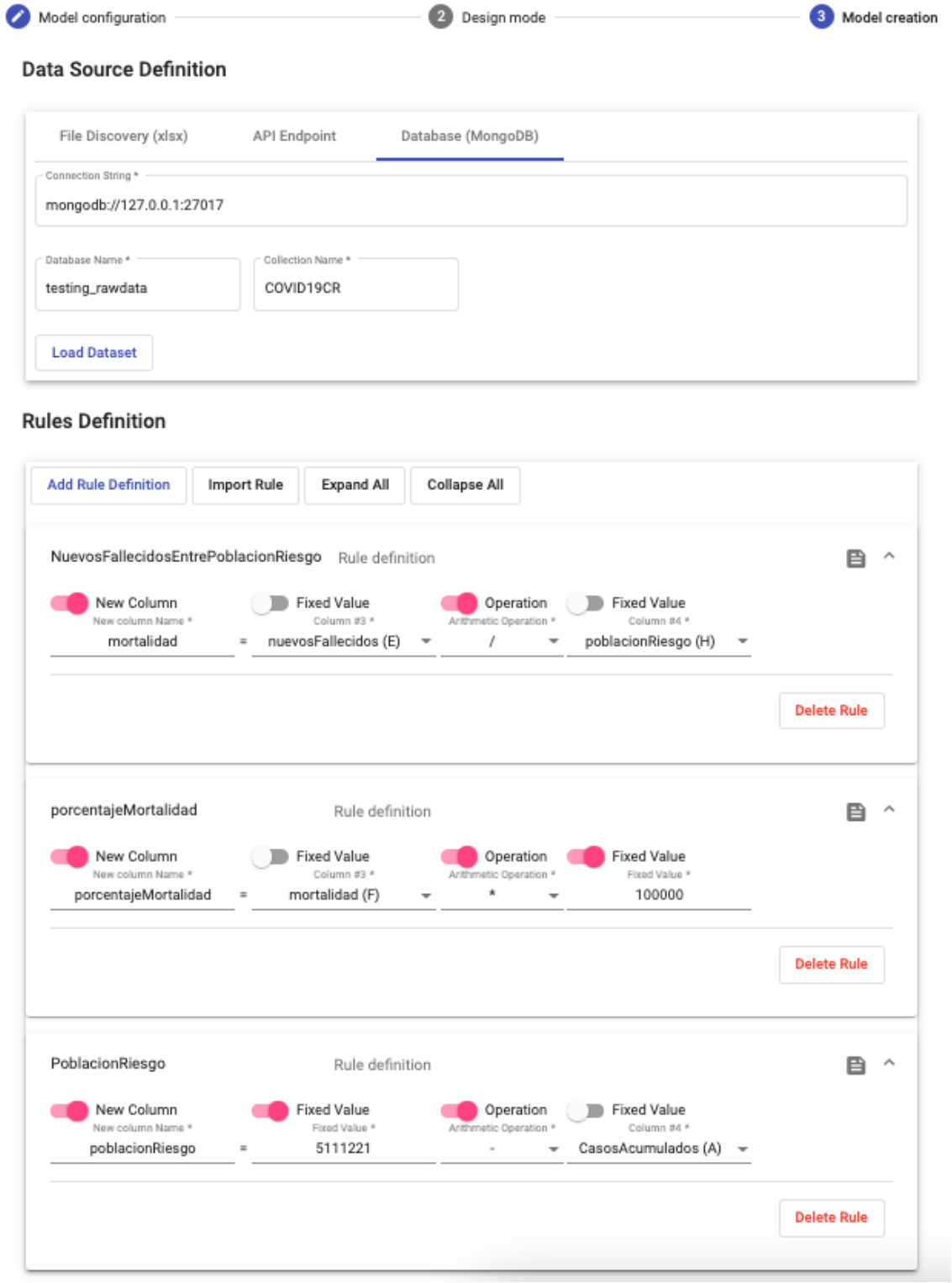
Anexo 3

Modelado de datos

Marca temporal	¿Que tan valioso cree que son los datos en una empresa?	¿Considera usted que el adecuado análisis de los datos es un aspecto estrictamente necesario para el éxito de una empresa a largo plazo?	¿Considera usted que el adecuado análisis de los datos es un aspecto estrictamente necesario para el éxito de una empresa a largo plazo?	¿Que base de datos utilizaría para almacenar datos que pueden cambiar con el tiempo y que necesitan estar bien estructurados?	Si tuviera que elegir un conjunto de datos sobre Costa Rica para crear nuevos modelos de datos. ¿Cual de las siguientes categorías elegiría?
2021/08/26 8:29:27 p. m. GMT-5	5	Sí	No conozco	Base de datos relaciones	Empleo
2021/09/06 3:52:22 p. m. GMT-5	5	Sí	SQL server, ERWin	Base de datos relaciones	Educación
2021/09/06 3:53:52 p. m. GMT-5	5	Sí	Keras	Base de datos no relaciones	Educación
2021/09/06 3:59:36 p. m. GMT-5	5	Sí	MySQL Workbench	Base de datos relaciones	Empleo
2021/09/06 9:49:40 p. m. GMT-5	5	Sí	NA	Base de datos no relaciones	COVID19

Anexo 4

Modelo completo creado con el modo de diseño asistido.



Dataset Preview

CasosAcumulados (A)	CasosporDia (B)	Dia (C)	Fecha (D)	nuevosF
2	2	1	06/03/2020	0
10	3	3	08/03/2020	0
12	2	4	09/03/2020	0
7	5	2	07/03/2020	0
m	1	5	10/03/2020	0
23	1	7	12/03/2020	0
22	9	6	11/03/2020	0

Anexo 5

Plan de pruebas

Ambiente de modelado de datos como servicio para facilitar el proceso de creación de modelos de datos en una organización

Propósito **80**

Responsabilidades **80**

Lineamientos para la ejecución de las pruebas **80**

Lineamientos para la creación de las pruebas **80**

Pruebas unitarias: 80

Pruebas de integración: 81

Pruebas prueba de extremo a extremo 81

Alcance de las pruebas **81**

Pruebas unitarias: 81

Pruebas de integración 81

Pruebas prueba de extremo a extremo 81

Propósito

El presente plan de pruebas, sirve como una guía para la creación y ejecución de pruebas para el producto de software desarrollado.

Responsabilidades

Responsabilidades de los desarrolladores:

- Crear pruebas unitarias para todo código nuevo desarrollado.
- Crear pruebas de integración para cada conjunto de tareas (métodos) comunes desarrolladas.
- Crear pruebas de extremo a extremo para cada funcionalidad nueva desarrollada.
- Actualizar las pruebas existentes que se afecten por los nuevos cambios en el código.

Lineamientos para la ejecución de las pruebas

A este punto de la herramienta, todas las pruebas deben ser ejecutadas en una herramienta de integración continua. De modo que cada vez que se integren nuevos cambios, se ejecuten todas las pruebas y se brinde la retroalimentación al equipo de desarrollo de manera efectiva.

Lineamientos para la creación de las pruebas

Las pruebas son creadas por el desarrollador que escribe el código correspondiente. Si durante el desarrollo se afecta una prueba ya existente, la prueba afectada es actualizada por el desarrollador que trabaja en el cambio.

- **Pruebas unitarias:**
  - Para el código en Angular, las pruebas deben ser agregadas en los archivos “specs.ts”.
  - Para el código Python, las pruebas deben ser agregadas en un proyecto nuevo con la misma estructura del proyecto principal que está siendo probado. Este proyecto nuevo debe existir en el mismo repositorio de toda la solución.

- **Pruebas de integración:**
    - Para el código en Angular, las pruebas deben ser agregadas en los archivos “specs.ts”, al igual que las pruebas unitarias.
    - Para el código Python, las pruebas deben ser agregadas en un proyecto nuevo con la misma estructura del proyecto principal que está siendo probado. Este proyecto nuevo debe existir en el mismo repositorio de toda la solución. Todo esto, igual que las pruebas unitarias.

- **Pruebas prueba de extremo a extremo:**

    Para estas pruebas, se utiliza el framework de pruebas de extremo a extremo llamado Cypress. Las pruebas deben ser agregadas al respectivo directorio de acuerdo a la especificación de este framework.

Alcance de las pruebas

- **Pruebas unitarias:**
    - Las pruebas unitarias deben probar partes individuales del código.

- **Pruebas de integración:**
    - Las pruebas de integración deben probar un conjunto de tareas funcionando de manera conjunta. Cada tarea es probada individualmente en las pruebas unitarias.

- **Pruebas prueba de extremo a extremo:**
    - Las pruebas de extremo a extremo deben verificar la correcta funcionalidad de una característica del sistema de principio a fin.

Anexo 6

Plan de mantenimiento

Ambiente de modelado de datos como servicio para facilitar el proceso de creación de modelos de datos en una organización

Propósito **84**

Responsabilidades **84**

Lineamientos para la ejecución de actualizaciones **84**

Lineamientos para la gestión de errores en el sistema **84**

Lineamientos para la gestión de nuevos requerimientos **85**

Propósito

El presente plan de mantenimiento sirve para mostrar la estrategia a seguir para el mantenimiento al sistema y mantenerlo actualizado.

Responsabilidades

Responsabilidades del equipo de desarrollo o persona a cargo del mantenimiento del sistema:

- Dar seguimiento a las herramientas utilizadas para conocer sobre eventuales vulnerabilidades detectadas.
- Corregir vulnerabilidades detectadas según corresponda en cada caso.
- Dar mantenimiento a los servidores utilizados a través de su gestión en la nube preferida.
- Actualizar software o infraestructura necesaria de acuerdo a los cambios en la industria.

Lineamientos para la ejecución de actualizaciones

Lo siguiente se logra a través del uso de contenedores con un orquestador.

- Utilizar estrategias que permitan actualizar el software con cero tiempo de inactividad, es decir, el servicio sigue funcionando en todo momento al liberar una actualización.
- Utilizar estrategias que permitan ejecutar con cero tiempo de inactividad al realizar cambios en infraestructura.
- Una vez que la actualización se complete, notificar a los usuarios acerca de los nuevos cambios.

Lineamientos para la gestión de errores en el sistema

- El desarrollador realiza la corrección del error correspondiente.
- Para cualquier error corregido, se debe agregar las pruebas correspondientes de acuerdo a lo establecido en el plan de pruebas.
- Para liberar la corrección del error, se debe actualizar el sistema siguiendo los lineamientos para actualizaciones.

Lineamientos para la gestión de nuevos requerimientos

- El desarrollador se asegura de comprender correctamente el requerimiento.
- Para todo nuevo requerimiento, se debe agregar las pruebas correspondientes según lo establecido en el plan de pruebas.
- Para liberar el nuevo requerimiento, se debe actualizar el sistema siguiendo los lineamientos para actualizaciones.