



Optimizar la colaboración y la toma de decisiones



## Documento de Análisis y Diseño

---

Para: **Adolfo Riebeling**

**Fecha:**  
**05 de diciembre de 2022**

Versión 2.0

---



## Contenido

1. Objetivo y Alcance del Ciclo de Negocio.....	3
2. Proceso de Negocio.....	3
3. Definición de Roles.....	10
4. Reportes .....	10
5. Catálogos .....	10
6. Workflows .....	10
7. Historial de revisiones.....	12
8. Firmas.....	12

## 2. Objetivo y Alcance del Ciclo de Negocio

### Objetivo

Desarrollar una aplicación web para agilizar los procesos de limpieza, enriquecimiento y preprocesamiento de los datos de los usuarios.

### Alcance

Se realizará una aplicación web mediante los lenguajes de programación ReactJS (v. 18.2.0), Python (v. 3.10.4) y una base de datos NoSQL, MongoDB (v. 6.0). El lenguaje ReactJS será utilizado para generar las ventanas de la aplicación web, además de formularios para la aplicación de los flujos, tablas de los datos del usuario y gráficas para cuantificar la actividad del cliente en la aplicación. El lenguaje Python será utilizado para la aplicación de diferentes funciones que se pueden resumir en *limpieza, enriquecimiento y preprocesamiento de los datos*. Por último, todos los datos de los usuarios, archivos cargados, flujos, historial, entre otros, serán guardados en una base de datos de MongoDB.

Se busca que con esta aplicación web los usuarios agilicen el proceso de tratamiento y validación de los datos, además de que puedan tener un repositorio de sus datos, sus flujos y un historial donde se vea reflejada toda su actividad mensual.

A continuación, se presenta un diagrama de la arquitectura que se tendrá en el DataHub.

## 3. Proceso de Negocio

### Narrativa a Alto Nivel del Ciclo de Negocio As Is

Actualmente las prácticas de Business Intelligence (BI) o Machine Learning (ML) han crecido de manera exponencial a lo largo de todas las empresas, pues ambas aportan información valiosa para la misma. Sin embargo, el éxito de un análisis de BI o el buen desempeño de los modelos de ML dependen ampliamente de la cantidad y la calidad de los datos, si alguna de estas características falla, nos podríamos enfrentar a un análisis sesgado o a un modelo pobremente entrenado.

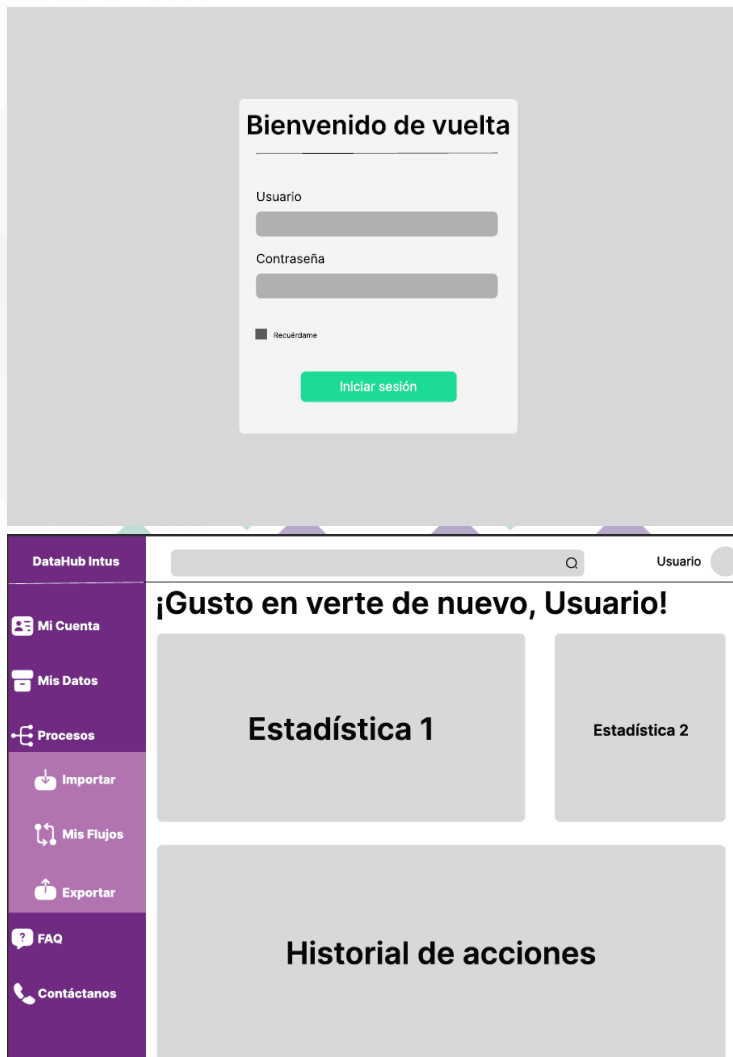
Con el desarrollo de este DataHub atacaríamos una de las principales debilidades de estas dos ramas, la calidad de los datos. Esto debido a que los usuarios tendrán la libertad de acceder a nuestra plataforma, cargar sus datos, establecer flujos para el tratamiento de sus

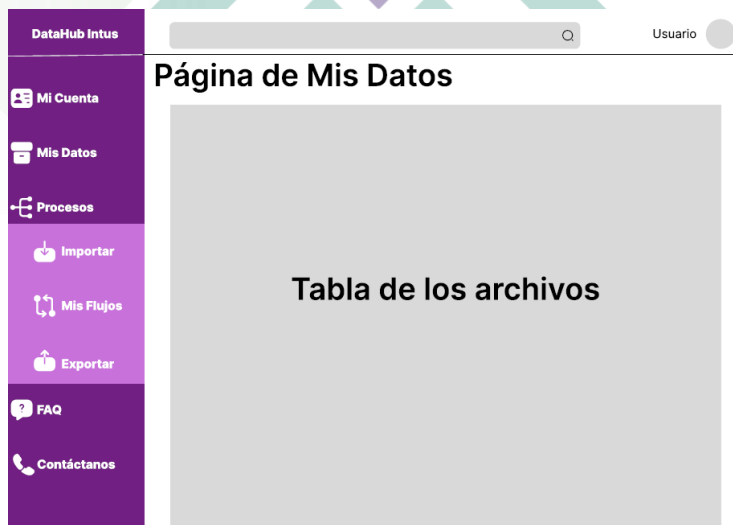
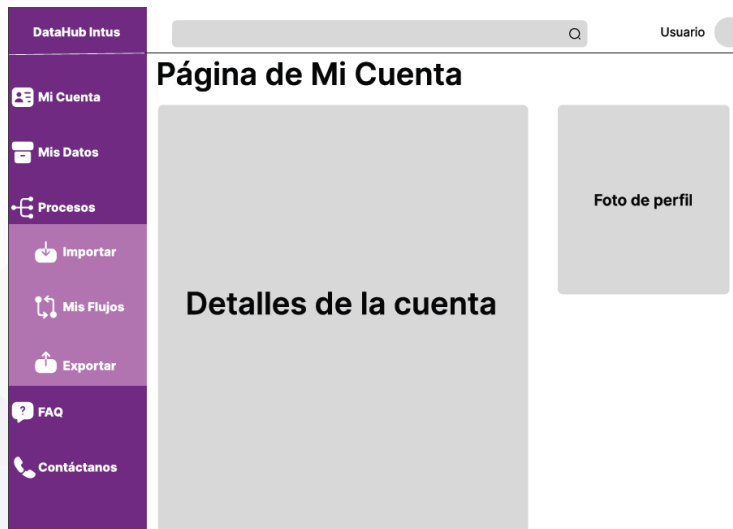
datos y exportarlos a una base de datos o descargarlos en el formato que cubra sus necesidades.


#### Datos generales

A continuación, se presentan diversas imágenes que corresponden a la primera aproximación al diseño final del DataHub.

Comentado [CA1]: Agregar las imágenes realizadas en Figma.








DataHub Intus  Usuario 


## Página de Importar


**Selección de archivo  
Nombrar archivo**


 Importar


 Mis Flujos

 Exportar


 FAQ

 Contáctanos





DataHub Intus  Usuario 


## Página de Mis Flujos





**Tabla de Flujos  
Información de Flujos  
Aplicación a Base de Datos**

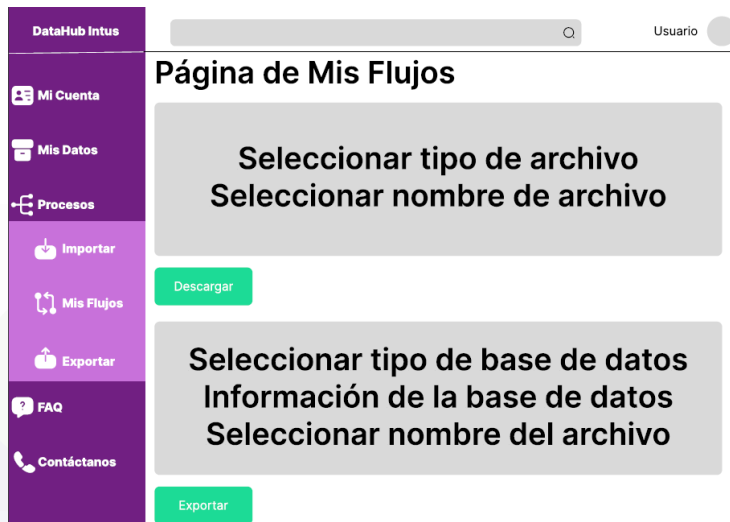
 Importar

 Mis Flujos

 Exportar

 FAQ

 Contáctanos



### 3.2.1. Página de Login

Sección desde la cual el usuario podrá ingresar a la aplicación web. Es necesario que previamente el usuario esté dado de alta en la base de datos de MongoDB para que sus datos puedan ser recolectados, es importante mencionar que en esta página también se generará el token de autenticación del usuario.

### 3.2.2. Página de Bienvenida

Sección principal de la aplicación web que consta de dos barras de navegación, una de ellas ubicada en la parte izquierda y la otra en la parte de arriba. Aquí el usuario podrá ver un resumen de sus bases de datos cargadas al sistema, los flujos que más ha utilizado y su actividad mensual en el DataHub.

### 3.2.3. Repositorio

Sección en la cual el usuario podrá encontrar sus orígenes de información como bases de datos o archivos CSV, TXT, XLSX, JSON, entre otros, que tiene registrados en el DataHub. Se le dará acceso al usuario de que cargue más bases de datos y que edite o elimine las ya existentes. Igualmente, podrá ejecutar sus flujos, acceder a sus perfilados de datos y los resultados de la aplicación de sus flujos.

### 3.2.4. Página de Importar

Sección en la que el usuario podrá subir nuevas fuentes de datos al DataHub, así como seleccionar un nombre clave para su proyecto y diversas opciones para la selección de los datos que quiera tratar.

### 3.2.5. Página de Mis Flujos

Sección donde el usuario podrá incluir reglas de limpieza, enriquecimiento y preprocesamiento de sus datos:

- **Limpieza:** Se refiere a la depuración de datos erróneos, caracteres especiales o datos irrelevantes que su única finalidad es generar ruido en nuestra base de datos.
- **Enriquecimiento:** Se entiende como el proceso de extraer nuevos datos relevantes a partir de los ya existentes en una fuente de datos.
- **Preprocesamiento:** Es el proceso en el cual se les da un formato específico a los datos para que puedan ser representados de buena manera por un modelo de ML o por un analista de BI.

Actualmente se cuentan con más de 30 reglas especializadas en los procesos mencionados y definidos anteriormente. Estas reglas se dividen en dos tipos, **predefinidas** y **personalizables**, las primeras de ellas son reglas encargadas de realizar una acción ya establecida como eliminar caracteres especiales, separar textos u obtener el tipo de dato de cada columna; mientras que las segundas tienen la capacidad de recibir parámetros del usuario para que se adaptan a lo que este está buscando. Cada conjunto de dos o más reglas se conoce como **flujo**, estos serán almacenados en esta misma sección y podrán ser manipulados por el usuario, además de darles un nombre para su fácil identificación. A continuación, se muestra el ejemplo de un flujo:

PROCESAR GOOGLE FORMS DE CLIENTES		
Orden	Regla	Descripción
1	Carga de fuente de datos.	Se comienza cargando la fuente de datos a utilizar.
2	Tipo de dato.	Se extrae el tipo de dato de cada una de las columnas de la fuente de datos.



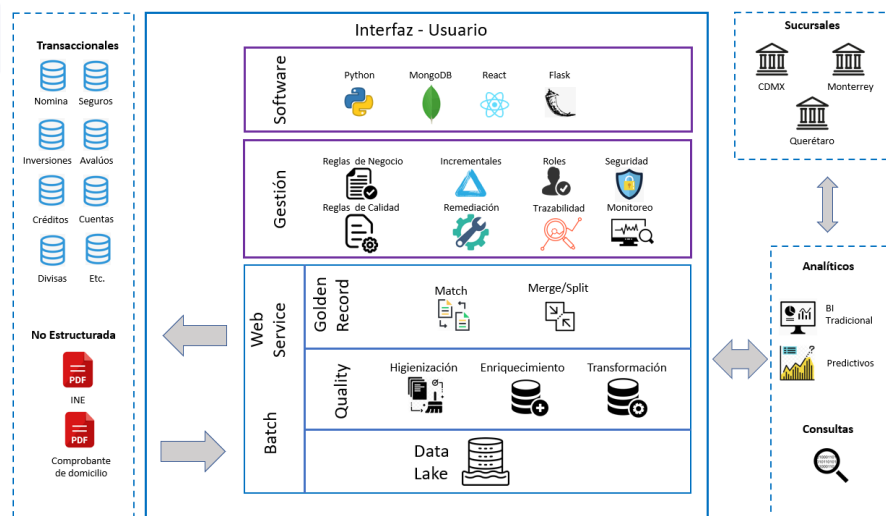
3	Eliminación Regexp.	Se eliminan los caracteres especiales especificados por el usuario que contiene la fuente de datos.
4	Eliminación HTML.	Se eliminan las etiquetas HTML que contiene la fuente de datos.
5	Comparación con catálogo.	Se compara el código postal con un catálogo de SEPOMEX.
6	Separación de texto.	Se separan nombre, apellido paterno y apellido materno en nuevas columnas.
7	Guardado de la fuente.	Se guarda la nueva fuente de datos ya tratada.

### 3.2.6. Página de Exportar

Sección desde la cual el usuario exportará los datos ya tratados a una base de datos externa (RDS, MySQL, etc.) o los descargará en el formato que más se acomode (CSV, XLSX, TXT, etc.).

### 3.2.7. Diagrama de Arquitectura

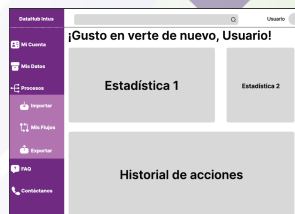
Estamos contemplando desarrollar en las siguientes tecnologías:



#### 4. Definición de Roles

Rol	Función	Permisos
<b>Cliente</b>	Hacer uso del DataHub e interactuar con la interfaz del mismo.	Lectura e interacción.
<b>Administrador de proyectos.</b>	Llevar el seguimiento del proyecto	Lectura.
<b>Líder del proyecto.</b>	Revisar el desarrollo del proyecto.	Modificaciones al código, lectura e interacción.
<b>Desarrolladores</b>	Realizar la interfaz del DataHub y en caso de ser necesario agregar o modificar funciones dentro del DataHub.	Modificaciones al código, lectura e interacción.

#### 5. Reportes



Citando la sección 2.2. aquí podremos observar un reporte de lo que el usuario ha estado realizando en la aplicación WEB del DataHub, como por ejemplo las estadísticas de sus archivos, o el historial de acciones que ha realizado, etc., es preciso mencionar que este reporte podría tener mayor o menor cantidad de resúmenes, esto dependiendo de la cantidad de archivos que el usuario procese en el DataHub.

#### 6. Catálogos

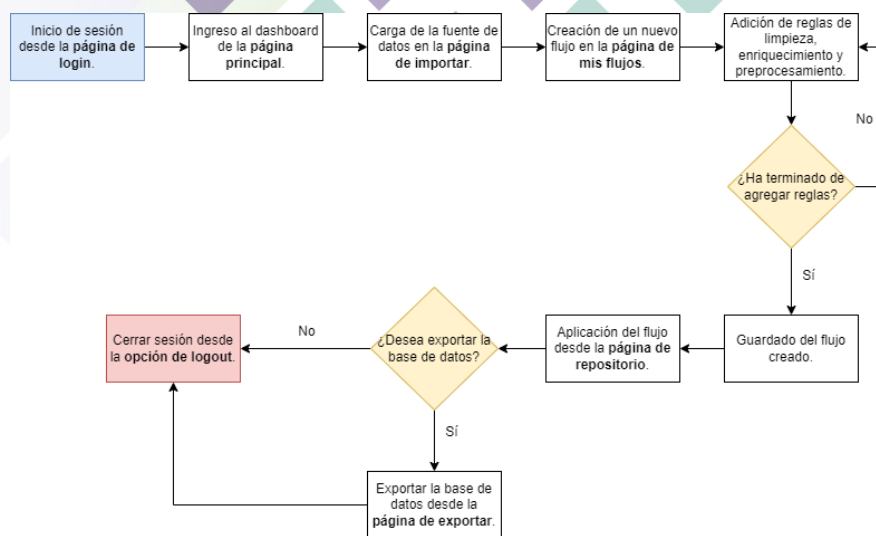
El inventario utilizado será para comparar la información que se tiene con datos reales (como por ejemplo los de SEPOMEX) para de esta manera tener información veraz y confiable.

#### 7. Workflows

Partiendo de que el usuario será quien interactúe con la aplicación WEB nuestros flujos de trabajo comienzan desde que ingresa a ella. A continuación, se enumera la manera ideal de

como seria el flujo, sin embargo, en ciertos puntos pueden omitirse o que se vuelva iterativo:

1. Darse de alta en la aplicación WEB.
2. Ingresar a la aplicación WEB.
3. Tablero donde se visualiza el resumen de interacción que ha tenido en la aplicación.
4. Visualización de los detalles de su cuenta.
5. Archivos ya cargados.
6. Importar archivos.
7. Tabla de flujos del DataHub.
8. Exportar.
9. Salir de la aplicación WEB.



## 8. Historial de revisiones

Inicio	Versión	Descripción	Autor
29/11/2022	V1	Primer documento sometido a revisión.	Carlos Ávila Joseph Martínez
05/12/2022	V2	Segundo documento sometido a revisión.	Carlos Ávila Joseph Martínez

## 9. Firmas

Realizó

Autorizó

Autorizó

Carlos Ávila, Joseph Martínez