

En búsqueda de la información

**Breve descripción:**

Este componente formativo aborda diferentes metodologías para desarrollar proyectos de ciencia de datos, apropiarse de diversas fuentes de información necesarias para la búsqueda eficiente de datos. Con su estudio responsable, el aprendiz podrá identificar algunos formatos del manejo de datos y, finalmente, sabrá desplegar una herramienta tecnológica basada en Python para la preparación, limpieza y transformación de datos.

**Septiembre 2023**

Tabla de contenido

[Introducción 1](#_Toc146897071)

[1. Metodologías de análisis de datos 3](#_Toc146897072)

[2. Fuentes de datos 11](#_Toc146897076)

[2.1. Tipos de bases de datos 12](#_Toc146897077)

[2.2. Tipos de archivos 17](#_Toc146897080)

[3. Infraestructura tecnológica 24](#_Toc146897084)

[3.1. Instalación 25](#_Toc146897086)

[3.2. Puesta en Marcha 26](#_Toc146897087)

[3.3. Instalación e importación de librerías 27](#_Toc146897088)

[4. Preparación de los datos 30](#_Toc146897089)

[Lectura de archivos en diferentes formatos 31](#_Toc146897090)

[Síntesis 37](#_Toc146897091)

[Material complementario 38](#_Toc146897092)

[Glosario 39](#_Toc146897093)

[Referencias bibliográficas 40](#_Toc146897094)

[Créditos 41](#_Toc146897095)

Introducción

Se da la bienvenida a este componente formativo denominado “**En búsqueda de la información**”. Comience el recorrido, explorando el video que se propone a continuación:

1. Búsqueda de la información



[Enlace de reproducción del video](https://youtu.be/wjYcgzSs-pc?si=fpTx3kU-IkUFdQMd)

|  |
| --- |
| **Síntesis del video: Búsqueda de la información** |
| Para desarrollar un análisis de datos efectivo, que arroje resultados significativos para la organización, es fundamental realizar un procedimiento inicial que permita buscar, jerarquizar y organizar los datos necesarios para la toma de decisiones y la formulación de estrategias que lleven al alcance de las metas y objetivos propuestos para el negocio.  Teniendo en cuenta lo anterior, este espacio de formación esta encamina a reconocer, en primera instancia, las diferentes metodologías a utilizar para la implementación de proyecto de analítica, entre las que se encuentran ASUM-DM (“Analytics Solutions Unified Method”), KDD (“Knowledge Discovery in Databases”), SEMMA (“Sample, Explore, Modify, Model, Assess”). Para el objetivo que le atañe a este curso se profundizará en la metodología CRISP-DM (“Cross-Industry Standard Process for Data Mining”), pues esta metodología resulta fundamental para llevar, de manera organizada, cada una de las etapas por la cuales debe pasar un proyecto de análisis.  Por otro lado, se reconocerán las principales fuentes de datos las cuales permiten obtener la información necesaria para el análisis. Se identificarán sus tipos y, mediante ejemplos, se reconocerá la mejor manera de utilizarlas.  De la misma forma, el desarrollo del presente escenario de formación le permitirá conocer los diferentes formatos en los que se puede manejar la información para realizar el análisis de datos; el uso de dichos formatos depende de los requerimientos y las necesidades organizacionales.  Finalmente, identificará cuál es el proceso para la instalación y el manejo de la herramienta **Anaconda**, que es una distribución del lenguaje **Python**, fundamental para la preparación, limpieza y transformación de los datos. |

# Metodologías de análisis de datos

Existen diferentes metodologías utilizadas para el análisis de los datos, para conocerlas, revise con atención la información que se presenta enseguida:

* **ASUM-DM**. (“Analytics Solutions Unified Method”): dicha metodología se enfoca en realizar implementaciones ágiles en temas de analítica de datos y minería de datos. A su vez, está basada en CRISP-DM, en la cual se dictan las especificaciones para llevar de manera ordenada y jerárquica tareas específicas para el análisis de los datos.
* **KDD**. (“Knowledge Discovery in Databases”): la metodología se basa en la evaluación e interpretación de patrones y modelos para tomar decisiones asertivas para el análisis. Dichos patrones deben caracterizarse por ser muy útiles y entendibles.
* **SEMMA**. (“Sample, Explore, Modify, Model, Assess”): la cual permite la selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio que hasta ahora han sido desconocidos.
* **CRISP-DM**. (“Cross-Industry Standard Process for Data Mining”): la cual permite llevar de manera organizada cada una de las etapas por la cuales debe pasar un proyecto de análisis de datos, incluyendo la limpieza y la trasformación de datos. Esta metodología facilita la creación de estrategias y el establecimiento de acciones necesarias para el cumplimiento de los objetivos propuestos en el negocio.

Teniendo en cuenta que la metodología que se debe trabajar en este curso es CRISP-DM (“Cross-Industry Standard Process for Data Mining”), es importante identificar los pasos para su implementación; el primer paso tiene que ver con identificar requerimientos del cliente o del negocio hacia el cual se va a hacer el análisis. Es importante definir las metas y objetivos de la organización, a partir de allí iniciar con la exploración de los datos y su categorización.

Posteriormente, se inicia con el proceso de preparación de los datos, aquí se crea el conjunto definitivo de los registros a los que se les va a realizar el análisis, se seleccionará el modelo a implementar para obtener los resultados esperados que serán sometidos a una etapa de evaluación; es importante tener en cuenta que todo este proceso es iterativo hasta perfeccionar cada fase y obtener los mejores resultados.

Para comprender mejor dicha metodología, se debe revisar con atención las etapas que se deben seguir para su correcta implementación:

### Etapa 1: Fase Inicial

Como ya se mencionó antes, lo primero que se debe hacer es el levantamiento de la información, esto permitirá evidenciar qué tan preparada está la organización para iniciarse en un proyecto de analítica de datos; dentro de esta etapa se identifican factores como:

* Tecnologías utilizadas por la organización.
* Las fuentes de datos con los que cuenta la organización.
* Cuáles son los medios para acceder a la información.
* Qué volumen de información maneja la organización.
* Con qué instalaciones y con qué infraestructura cuenta la organización.

Por otro lado, es fundamental identificar los roles que, dentro de la organización, son necesarios para desarrollar a cabo el proyecto de análisis. Conózcalos a continuación:

* **Patrocinador del proyecto**. Es la persona dentro de la organización, de nivel superior, encargada de tomar las decisiones importantes y responsable de que el proyecto tenga los insumos tecnológicos y humanos requeridos.
* **Arquitecto empresarial**. Su responsabilidad se basa en mantener todo el inventario, a nivel tecnológico, al día y acorde a las metas propuestas por la empresa.
* **Arquitecto de datos**. Se encarga de garantizar la infraestructura necesaria para soportar los proyectos de ciencia de datos.
* **Científico de datos**. Encargado de los datos, tiene un amplio dominio estadístico y de herramientas tecnológicas para el procesamiento y análisis de datos.
* **Analista de negocio**. Personal de la empresa que domina muy bien los objetivos y metas, se encarga de las estrategias empresariales y en proyectos de análisis de datos debe trabajar de la mano con el científico de datos.
* **Gerente de datos y analítica**. Conoce todas las etapas y responsables del proyecto, se encarga de que todas las áreas de la empresa engranen y la información fluya de la mejor manera.

### Etapa 2: Comprender el negocio

Durante esta etapa es importante tener claridad frente a los objetivos del negocio, además es fundamental determinar cuáles son los requerimientos y las necesidades de la empresa, teniendo claros dichos aspectos se pueden empezar a visualizar posibles problemas que tiene la empresa y que se quieren resolver con la implementación de la analítica.

Para el desarrollo de esta etapa es importante tener en cuenta algunos aspectos, como:

* **Evaluación de la situación**. Evaluación que se fundamenta en los objetivos de la empresa o negocio. En dicha evaluación es importante incluir: inventario, requerimientos, riesgos y planes de contingencia, glosario, análisis de costo-beneficio, ruta de trabajo, reportes.
* **Recorrido histórico**.Para determinar los objetivos organizacionales se puede realizar un recuento histórico de negocio que permita establecer la situación actual, identificando debilidades y fortalezas, a partir de allí se pueden establecer los objetivos del análisis.
* **Estado actual o situación sincrónica**. Una vez se tiene claridad frente a los objetivos del negocio, se debe realizar una evaluación que determine el estado actual de la organización en aspectos como: la empresa cuenta con el personal necesario para completar el proyecto, con qué tipos de datos cuenta la organización para realizar el análisis, cuáles pueden ser los factores de riesgo con los que cuenta la organización, se disponen de planes de contingencia para cada factor de riesgo.
* **Identificación de áreas y de personal**. Es fundamental identificar el personal por área con el que cuenta la empresa, el conocimiento de negocio, las habilidades tecnológicas con las que cuentan cada uno de ellos. Por ejemplo, los administradores de bases de datos pueden ser útiles para realizar la extracción de datos. Además, es importante determinar los recursos informáticos y las herramientas de “software” de las que dispone la organización, como equipos, servidores, capacidades de procesamiento, bases de datos y “software” de extracción y análisis de datos.
* **Cronograma de actividades**. Es importante realizar un cronograma de actividades para el desarrollo del proyecto. Esto permitirá tener una visualización clara de cada una de las tareas con sus respectivos responsables. Además, cuando se trabaja con datos, es fundamental tener claridad sobre las políticas de protección de datos. Estas políticas suelen estar reguladas por leyes estatales y estatutos propios de la empresa. Por último, se debe analizar la existencia de limitaciones en cuanto a “hardware” y “software”. En caso de que existan, es crucial comunicar los requerimientos necesarios a las partes correspondientes.
* **Identificación de riesgos**. Es necesario identificar los riesgos visibles que pueden surgir durante el proceso. Por ejemplo, en el ámbito de los recursos humanos, podrían incluir el despido de personal, mientras que en el ámbito tecnológico, podrían incluirse fallas en equipos o herramientas cruciales para el proyecto. Por lo tanto, resulta fundamental contar con una ruta estructurada, documentada en un archivo maestro, que permita tomar medidas para abordar las posibles eventualidades que puedan surgir.
* **Comunicación eficaz**. Para tener una comunicación fluida entre todas las personas que hacen parte de la organización y las que trabajan directamente con el desarrollo de proyecto, es importante socializar un documento y realizar capacitaciones frente a la terminología utilizada.
* **Reportes**. Es necesario realizar un informe detallado que incluya los costos y beneficios que puede traer la implementación del proyecto. Es importante tener en cuenta tanto los costos directos como los indirectos en el proceso de recolección de los datos y en el despliegue del proyecto. Asimismo, es fundamental documentar los posibles beneficios que dicha implementación traerá a la organización una vez finalizada.
* **Claridad en los propósitos**. Para tener éxito en el proyecto, es importante tener claro el paso a paso: qué se va a realizar en cada una de las etapas del proceso. De igual manera, los involucrados y las responsabilidades deben estar bien definidos.
* **Beneficios de comprender el negocio**. Para el equipo de analítica es vital comprender el negocio completamente, el objetivo que se desea alcanzar, los criterios de éxito y demás datos relevantes para entender el negocio, con los reportes se genera un insumo fundamental que puede ser ejecutado en cualquier momento durante el desarrollo del proyecto.

### Etapa 3: Fuentes de información y descripción de los datos

Para llevar a cabo un análisis de datos exitoso, es necesario realizar una recolección de datos iniciales que permita determinar la consistencia de la información. Se espera que en dicha recolección sea posible describirlos, llevar a cabo todas las exploraciones planteadas y realizar la verificación de la calidad de los mismos.

En ese orden de ideas, existen diferentes fuentes de información, es decir, los diversos medios a través de los cuales se pueden adquirir los datos. Puede ser que la organización tenga sus datos en sus sistemas privados, de los cuales es propietaria, o también puede adquirirlos desde sistemas externos (ya sean de pago o gratuitos).

Dentro de los tipos de fuentes de información se tienen:

1. **Fuentes propias de datos**. Es todo de tipo de información privada, de la cual se es propietario y se tiene completo control, como, por ejemplo, sistemas de información, encuestas, movimientos transaccionales y demás datos relevantes de la compañía.
2. **Fuentes de datos externos**. En caso de que la empresa no posea cierto tipo de datos puede optar por la compra de estos, de igual manera puede hacer uso de datos abiertos disponibles gratuitamente en internet.

Después de realizar la exploración de los datos por medio de las diferentes fuentes, es importante empezar a realizar la descripción de cada colección analizada; generalmente, estos deben ser organizados en filas y columnas, las columnas pueden estar determinadas por dos tipos de datos, así:

1. **Datos categóricos**. Este tipo de datos hacen referencia a valores cualitativos como, por ejemplo, el género, color, marca, estado. Estos son valores que no se pueden cuantificar.
2. **Datos numéricos**. Representan un valor que se puede cuantificar de forma numérica y a los cuáles se les puede aplicar diversos métodos estadísticos, dentro de este tipo de datos podemos definir edad, salario, temperatura, etc. Al ser numéricos se pueden sumar, promediar y en general aplicar todas las operaciones matemáticas y estadísticas.

Después de determinar los datos categóricos y numéricos es fundamental establecer las medidas de tendencia central como la moda, media, mediana, varianzas, máximos y mínimos.

Dentro de esta etapa 3, también es necesario realizar una verificación detallada de la calidad de los datos, para esto es importante tener en cuenta aspectos como:

* ¿Los datos están completos?
* ¿Los datos son correctos?
* ¿Qué tipo de errores se pueden presentar durante el análisis?
* Hay valores que son nulos, ¿qué hacer si esto se presenta?
* Hay datos con formatos incorrectos, ¿qué hacer si esto se presenta?
* Hay datos duplicados, ¿qué hacer si esto se presenta?

Finalmente, es importante la generación de los informes que indiquen el estado en que se encontraron los datos. En este, se debe considerar el proceso que se llevó a cabo para la recolección de los datos, la descripción encontrada del estado de los datos y la descripción del proceso de exploración y calidad de los datos.

El informe debe incluir el número total de registros, la clasificación de cada columna en numérica y categórica, una breve descripción de lo que cada columna representa, se debe identificar cuáles columnas presentan datos vacíos, si existen valores duplicados, y qué columnas se encuentran en un formato incorrecto.

# Fuentes de datos

Es fundamental tener total claridad frente a este tema, pues el valor de la información que maneja la organización se determina también a partir de las fuentes de donde estas son extraídas. Como ya se mencionó anteriormente, las fuentes de datos pueden ser privadas o públicas. Respecto a las fuentes privadas, los datos se pueden obtener desde sistemas de información, bases de datos, hojas de cálculo, entre otros recursos en los cuales la organización almacena la información propia. En cuanto a las fuentes públicas, estas permiten obtener información desde internet de manera libre o mediante un registro de usuario.

A continuación, conozca algunas fuentes de datos:

* **Bases de datos**. “Software” utilizado para almacenar, consultar, modificar o eliminar la información directamente o desde sistemas externos como ERP o CRM.
* **ERP**. Son “Software” que se utilizan para administrar las operaciones de las empresas o negocios, como el manejo de clientes, cartera, contabilidad, recursos humanos, manufactura, entre otras.
* **CRM**. Sistema que se especializa en las relaciones con el cliente y busca capturar toda la información necesaria para fidelizarlo y establecer estrategias para captar otros nuevos clientes.
* **Formularios de consulta**. Utilizados para la aplicación de encuestas y la recolección de información masiva de un grupo objetivo, estos se comparten fácilmente por redes sociales y aplicaciones de mensajería.
* **Redes sociales**. Permiten la captura de información referente a estados de ánimo, publicaciones, preferencias, gustos y diferente tipo de información; para acceder a dicha información se necesitan una autenticación con usuario y contraseña.
* **IoT Internet de las cosas**. Gracias al internet de las cosas se puede capturar información automática, como las variables de clima, temperatura de un departamento o país, las capturas del tráfico y los estados de movilidad de una ciudad determinada.
* **Bases de datos gubernamentales**. Aquí se reportan todas las entidades que hacen parte del Estado; se puede encontrar información como, por ejemplo, resultados de pruebas ICFES, contrataciones, entre otros.

Actualmente se pueden encontrar fuentes de datos públicas de salud, economía, construcción, tendencias y casi que cualquier tipo de tema que se quiera analizar, por ende, se puede afirmar que los datos se han convertido en los activos más importantes para cualquier organización, pues ellos son la base para la toma de decisiones estratégicas, por tal motivo es necesario establecer protocolos y normas para su manejo, teniendo en cuenta criterios como confidencialidad, disponibilidad e integridad.

## Tipos de bases de datos

Todo tipo de dato que se obtiene diariamente dentro de la organización, por medio de empleados, sistemas de información, redes sociales, etc., son un insumo de gran importancia para el cumplimiento de los objetivos planteados por la misma. Por tanto, estos deben ser almacenados de manera correcta de modo que se puedan consultar, modificar o eliminar. Para tal almacenamiento se hace uso, generalmente, de dos tipos de bases de datos: las bases de datos relacionales y las no relacionales.

### Bases de datos relacionales

Este tipo de base de datos siguen una estructura estricta y se basan en tablas que almacenan la información organizada, de acuerdo con el tipo de dato.

A continuación, se mencionan los programas más utilizados para bases de datos relacionales:

1. **Microsoft SQL Server**. Es utilizado para trabajar conjuntos de datos y crear relaciones entre ellos, es desarrollado por la empresa Microsoft, para usos comerciales, por ende, requiere de licencia.
2. **DB2**. Base de datos desarrollada por IBM. Para la versión comercial requiere licencia.
3. **Postgree SQL**. Sistema de base de datos relacional de código abierto, orientado a objetos.
4. **MariaDB**. Igualmente, de código abierto utilizada para proyectos más pequeños que no requieren gran volumen y procesamiento de datos.
5. **MySQL**. Es una de las bases de datos más utilizadas en el mundo, se conecta con múltiples sistemas de información y sistemas operativos, maneja una versión gratis y versiones comerciales.

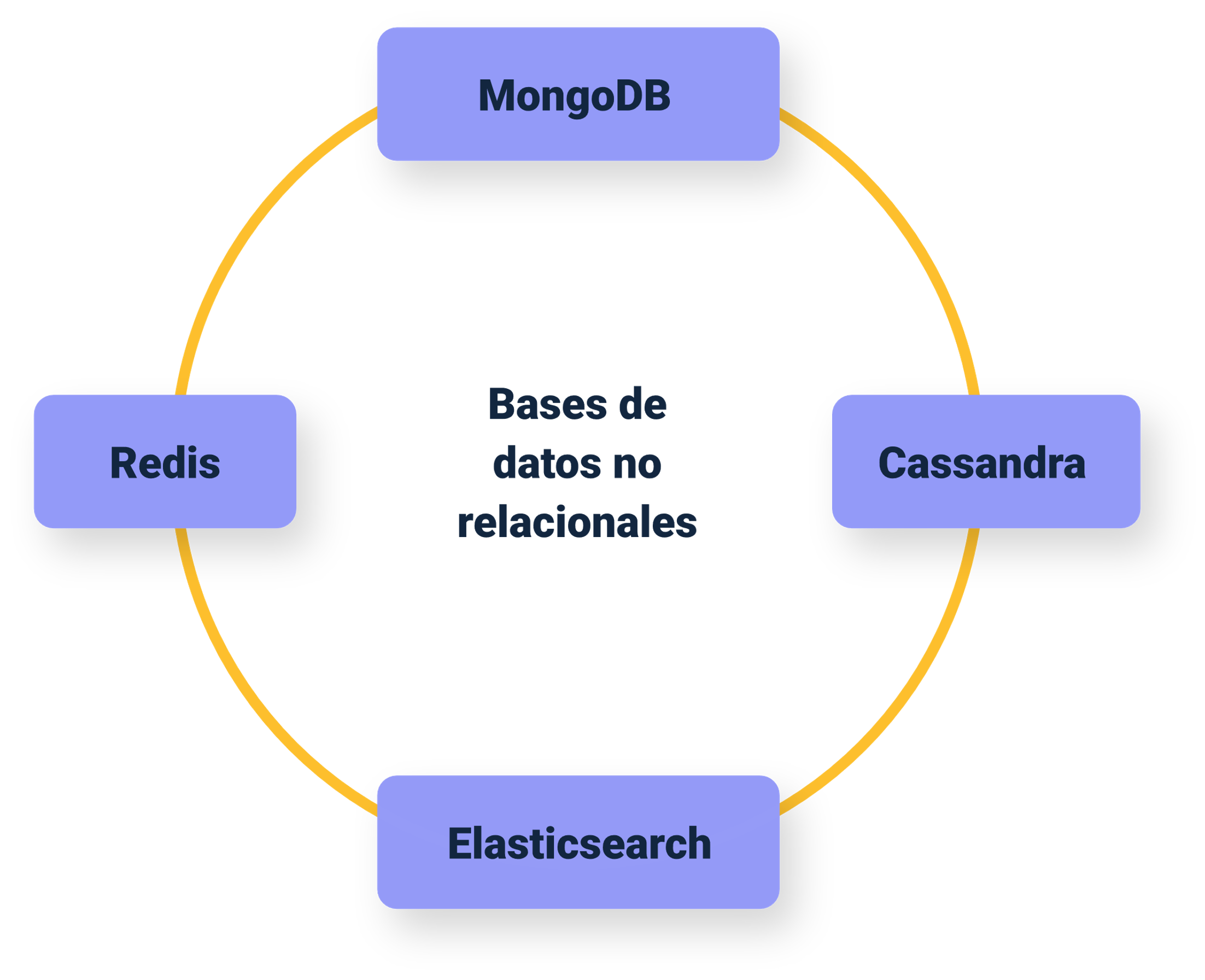
Las bases de datos relacionales hacen referencia a una base de datos cuya característica es almacenar y proporcionar acceso a puntos de datos relacionados entre sí, es una forma directa de representar datos en tablas.

Para entender mejor en qué consisten las bases de datos relacionales, se le recomienda revisar en detalle el PDF denominado **Bases de datos relacionales**, que se encuentra en la carpeta Anexos.

### Bases de datos no relacionales

También conocidas como NoSQL, este es un sistema que se caracteriza por no usar el lenguaje SQL para realizar consultas. A diferencia de las bases de datos relacionales, no utiliza tablas y la información no es almacenada como registros, lo hace mediante colecciones y documentos, son muy útiles para almacenar grandes volúmenes de información.

1. Bases de datos no relacionales

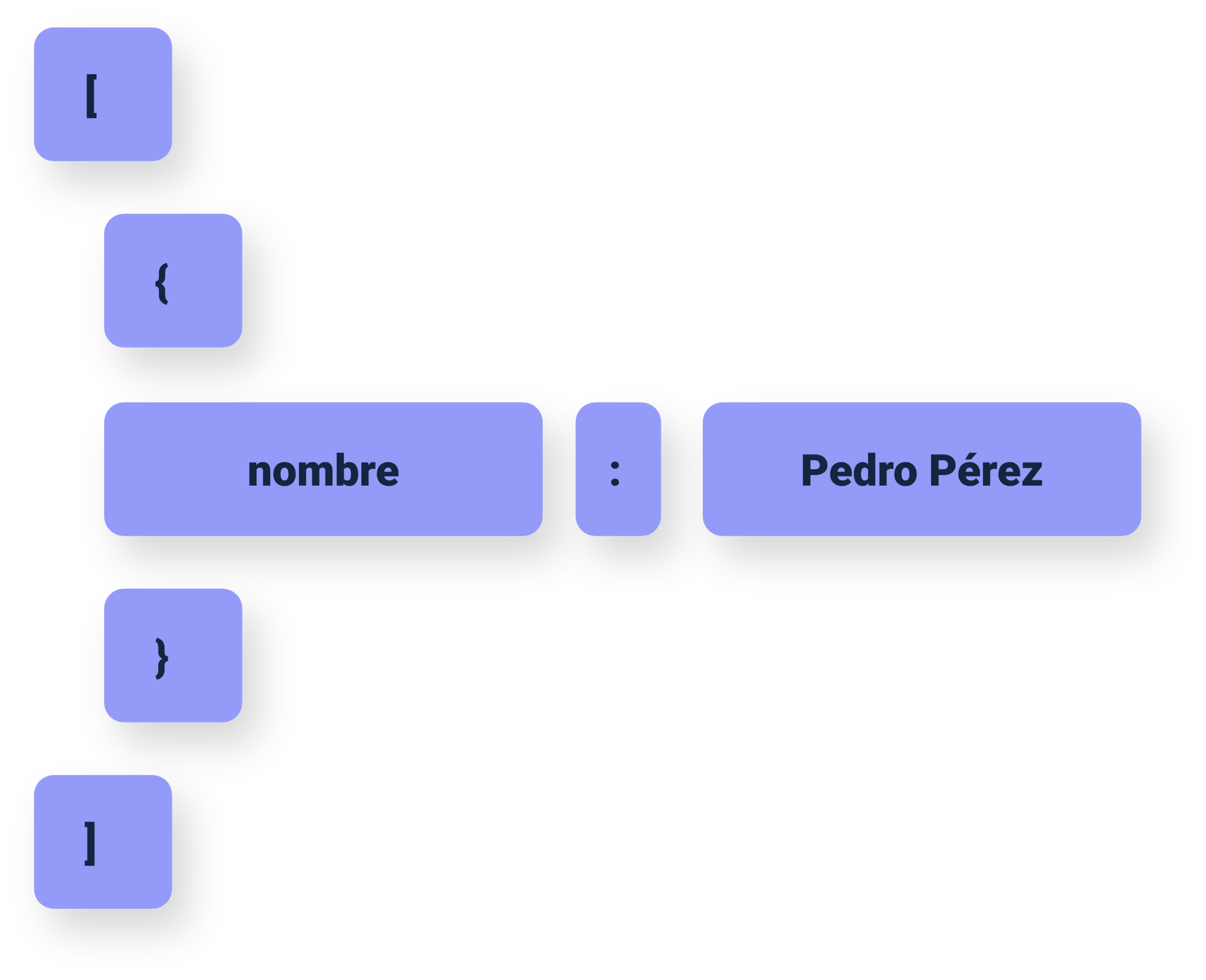


Las bases de datos no relacionales usan, entre otros, sistemas como MongoDB, Casandra, Redis y Elasticresearch, los cuales almacenan información en colecciones y/o documentos.

Dentro de las bases de datos no relacionales, la información es almacenada mediante corchetes que hacen referencia a una lista o colecciones y cada documento es almacenado dentro de llaves con la información, ya no en forma de filas y columnas sino de clave y valor. La clave o identificador se coloca a la izquierda de los dos puntos y el valor a la derecha, como se muestra a continuación:

En el ejemplo la clave es representado por el nombre y valor como Pedro Pérez:

1. Almacenamiento en bases de datos no relacionales



El almacenamiento en bases de datos no relacionales, se estructura: Llave – corchete – Clave o identificador - dos puntos (:) – Valor.

A continuación, se presenta un ejemplo de una colección que tiene cuatro documentos, con información relacionada a empleados de una empresa:

[

{

"Id\_Cliente":1001,

"nombre":"Pedro Perez",

"edad": 23,

"Ciudad":"Bogota",

"Cargo":"Gerente",

}

{

"Id\_Cliente":1002,

"nombre":"Maria Rojas",

"edad": 45,

"Ciudad":"Medellin",

"Cargo":"Vendedor",

}

{

"Id\_Cliente":1003,

"nombre":"Pedro Perez",

"edad": 28,

"Ciudad":"Barranquilla",

"Cargo":"Vendedor",

}

{

"Id\_Cliente":1004,

"nombre":"Pedro Perez",

"edad": 32,

"Ciudad":"Cali",

"Cargo":"Contador",

}

]

## Tipos de archivos

La información que es almacenada en las bases de datos, ya sean relacionales o no relacionales, puede ser extraída en diferentes tipos de archivo. Dependiendo del tipo de archivo, la información se visualiza de diversas maneras; son múltiples los tipos de archivos en los que se almacena la información. Para este caso se abarcarán algunos de los más comunes en análisis de datos:

### Archivos XLS, XLSX

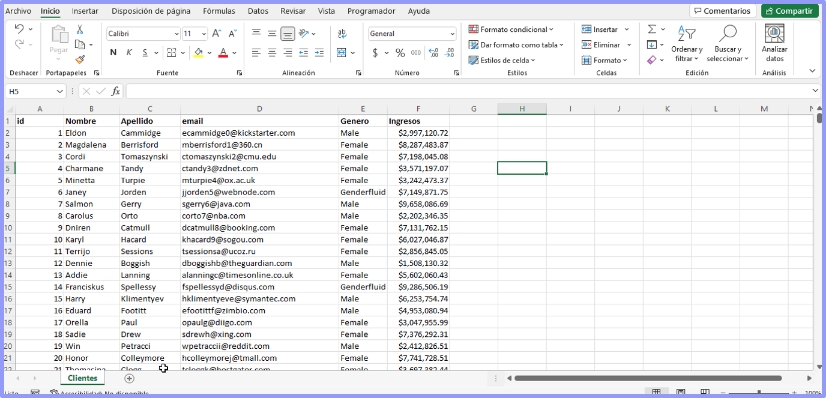
Este tipo de archivos se generan desde uno de los programas más conocidos en el mundo de la informática para herramientas ofimáticas como lo es Microsoft Excel, las extensiones de los archivos se generan con extensión XLS para versiones 2003 y XLSX para versiones desde 2007 en adelante. Este tipo de archivos, actualmente, se generan por múltiples sistemas de información: formularios en línea, office en línea y por muchas otras aplicaciones.

Conozca, enseguida, las características más importantes de estos archivos:

* Utilizan filas y columnas.
* Las filas se identifican con números consecutivos.
* Las columnas se identifican con letras en orden alfabético.
* La intersección entre una columna y una fila se denomina celda.

A continuación, se presenta un ejemplo de un archivo que relaciona un listado de clientes; las columnas se utilizan para identificar el tipo de datos y las filas para ingresar cada uno de los registros que pertenecen a esa colección de datos:

1. Archivo en formato XLSX



### Archivos CSV

No es un formato tan popular con el cual se esté familiarizados, pues no se utiliza para tareas cotidianas como sí se hace con las hojas de cálculo de Excel u otro programa, para el manejo de archivos. Este es un tipo de archivo que maneja una estructura simple y almacena la información en forma de filas que corresponden a cada uno de los registros y los campos de esos registros están, a su vez, separadas normalmente por comas, aunque también se pueden encontrar separados por punto y coma o tabulaciones.

El CSV es un tipo de formato que se puede abrir casi en cualquier programa como, por ejemplo, un sencillo “bloc” de notas, aunque allí la compresión de los datos no es tan clara, por ende, en hojas de cálculo como Excel, Google Sheet, u otros programas compatibles con esta estructura, es un formato que no admite gráficos, tablas, colores o fuentes y está diseñado, simplemente, como un contendedor de información en la forma más simple.

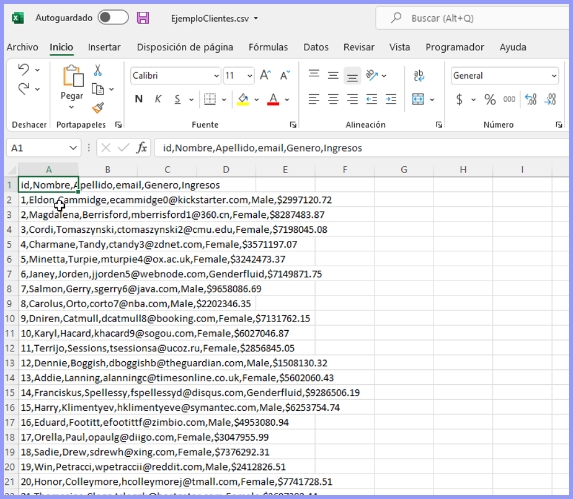
Pero, ¿por qué utilizar este tipo de formatos?

1. **Simplicidad**. La simpleza con la que se guarda la información permite almacenar grandes volúmenes de datos.
2. **Almacenaje amplio**. Al ser un formato ligero se utiliza como contenedor.
3. **Exportación de información**. Luego de ser usado como contenedor se puede realizar, desde allí, la exportación a sistemas de información, directamente de base de datos, o desde las mismas hojas de cálculo sin comprometer el rendimiento de los sistemas.

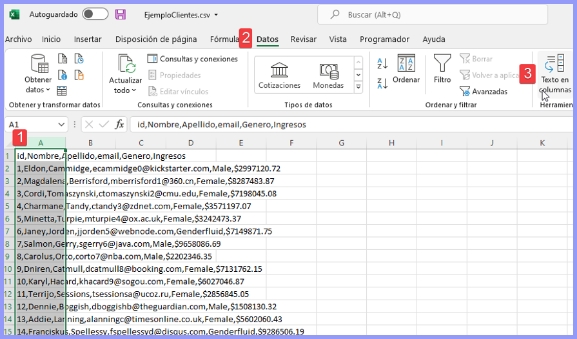
Este es uno de los formatos más utilizados para los proyectos de analítica de datos en los que se requiere procesar grandes volúmenes de información, por ende, desde la preparación de los datos es importante ir planeando el proceso para optimización de recursos, tanto en el almacenamiento como en el procesamiento, en ese sentido los archivos en este tipo de formato son altamente recomendados.

A continuación, conozca los pasos para convertir un archivo en formato CSV a formato XLSX:

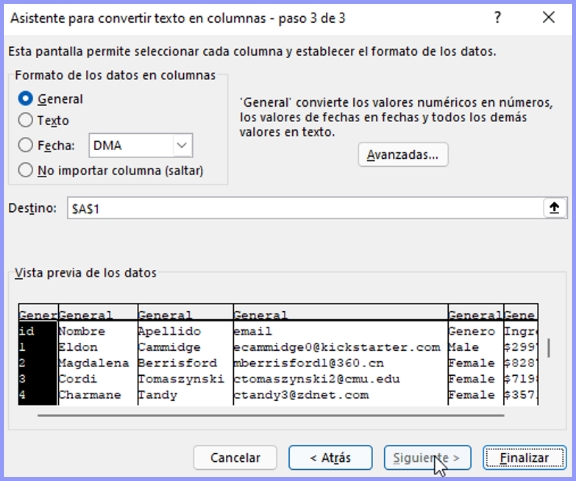
1. Desde la hoja de cálculo Excel, abrir el archivo formato CSV.



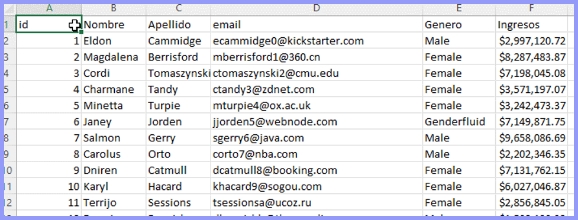
1. El objetivo es separar, dividendo por comas, cada uno de los registros para poder visualizarlos correctamente en filas y columnas. Para ello se selecciona toda la columna A, luego se selecciona la pestaña datos y se presiona texto en columnas.



1. Entonces, se deben separar los datos por comas; una vez seleccionado texto en columnas se abrirá el asistente en el cual se podrá elegir cómo separar los registros. Se deben seleccionar delimitados, posteriormente seleccionar que la separación se realizará por comas.



1. Para terminar, se selecciona finalizar y se podrán tener los registros separados en columnas.



### Archivos JSON

Más que un tipo de archivo es una estructura simple que se utiliza para el intercambio de información entre sitios web, cuando se realiza la exportación de este tipo de archivos se utilizan archivos de texto simple con extensión TXT o JSON. Este tipo de estructuras se puede obtener desde bases de datos no relacionales, cuando es necesario compartir grandes volúmenes de datos o también para el intercambio de información entre sistemas.

El siguiente es un ejemplo de cómo sería la representación de un documento que muestra la información de las ventas de un empleado:

{

“Código”: 1001,

“Nombre”: “Pedro Perez”,

“Género”: “M”

Ventas: 1200000

}

Para entender mejor la estructura identifique la equivalencia que tiene esta estructura en filas y columnas, configurándose cada uno de los valores de la izquierda en columnas y, los de la derecha, en cada uno de los registros o filas:

1. Equivalencia filas y columnas

| Código | Nombre | Género | Ventas |
| --- | --- | --- | --- |
| 1001 | Pedro Pérez | M | 12000000 |

# Infraestructura tecnológica

La herramienta utilizada en este curso para la lectura y manipulación de los archivos se basa en una distribución de Python llamada Anaconda, la cual es de uso gratuito, descargándola desde su sitio oficial.

Es importante tener en cuenta como todo “software” que se vaya a utilizar dentro de un equipo de cómputo, requiere de unas condiciones técnicas para que el aplicativo funcione correctamente, por eso a continuación se hará una descripción detallada de dichas consideraciones y se dará a conocer el paso a paso para la instalación y despliegue de las herramientas de “software”.

Para la implementación de una infraestructura tecnológica existe un concepto importante que tiene que ver con la unificación de diversas fuentes de información y herramientas de análisis, que optimizan las tareas de lectura, adecuación y procesamiento de los datos, a continuación, se reconocerá esta tecnología.

### Arquitectura Data Warehouse

Con los formatos y estructuras de los archivos plenamente identificados se puede pensar en la unificación de sistema propio de almacenamiento que permita, independiente de la estructura o formato, tener toda la información en única base de datos normalizada, esta base de datos estará preparada para recibir grandes volúmenes de información, actualizarse, modificarse y consultarse. A este tipo de arquitectura se le conoce como DATA WAREHOUSE.

¡**Nota importante**! continuando con la explicación de la infraestructura tecnológica y referente a lo que compete a este curso, se darán a conocer los pasos y consideraciones que se deben tener en cuenta para la instalación de la herramienta Anaconda. Es clave seguir, atentamente, las instrucciones para que la instalación sea exitosa.

## Instalación

Para empezar con la instalación de las herramientas tecnológicas que permita el despliegue del lenguaje de programación Python, es importante que se valide los requerimientos mínimos técnicos con los que debe contar su equipo ya sea portátil o de escritorio, estos son:

* Memoria RAM: 3GB.
* Disco Duro: mínimo 4 GB espacio disponible.
* Procesador: procesador de 32- o 64-bit.
* Sistema Operativo: Windows, Linux y Mac.

Posteriormente a la verificación de los requerimientos técnicos, se procede con la descarga del programa Anaconda, disponible en la página oficial.

Preste atención al siguiente video, donde se explica este procedimiento:

1. Instalación de Anaconda



[Enlace de reproducción del video](https://youtu.be/wSdQpgVSPvY?si=BxHiWMxwghQCdokY)

|  |
| --- |
| **Síntesis del video: Instalación de Anaconda** |
| En este video, el instructor SENA desarrolla una explicación detallada de pasos y acciones para la instalación correcta de Anaconda. Ofrece un ejemplo real de instalación en pantalla para que sea consultado por cualquier persona interesada que tenga acceso al video.  Siga el video con suma atención. |

## Puesta en Marcha

Es importante tener en cuenta que Anaconda maneja múltiples aplicativos, para este curso en particular se utilizará JupyterLab, la cual es una herramienta web que permite, mediante una interfaz gráfica, ejecutar líneas de comando del lenguaje de programación Python. Estas líneas de código se ejecutan en celdas, de manera independiente, lo que permite ir probando cada parte del programa facilitando la detección de errores pues al ejecutarse, línea a línea, se puede saber exactamente dónde se encuentra la dificultad.

Por otro lado, Python, gracias a sus aplicaciones y gran número de librerías, es el lenguaje de programación mayormente utilizado para proyectos de ciencia de datos. JupyterLab es una herramienta muy intuitiva y muy fácil de usar, lo que permite que el despliegue sea realmente sencillo.

Para empezar a trabajar en un proyecto de desarrollo se debe abrir la aplicación Jupyterlab; para realizar este procedimiento, se abre el programa Anaconda y se busca la aplicación Anaconda Navigator y, dentro de sus aplicaciones, se puede encontrar dicha herramienta.

Sepa cómo crear el entorno de trabajo; atienda la explicación y dé los pasos necesarios, los cuales se ofrecen en el siguiente video:

1. Puesta en marcha



[Enlace de reproducción del video](https://youtu.be/OanYK6mqBIo?si=iD8l9y3HfwZKCOBp)

|  |
| --- |
| **Síntesis del video: Puesta en marcha** |
| En este video, el instructor SENA desarrolla una explicación detallada de cómo crear el entorno de trabajo y de pasos y acciones para la puesta en marcha. Ofrece un ejemplo real del proceso, en pantalla, para que sea consultado por cualquier persona interesada que tenga acceso al video. |

## Instalación e importación de librerías

Hasta este punto, ya se conocen los procedimientos básicos para instalar las herramientas necesarias para realizar el análisis de datos; igualmente, ya se sabe cómo desplegar el entorno de trabajo JupyterLab, así como se han identificado cada uno de los elementos que la conforman. En este apartado se profundizará en cada una de las librerías necesarias para trabajar proyectos de ciencia de datos, puntualmente las que ayudarán en todo el proceso de limpieza y transformación de los datos.

Estos son algunos aspectos clave que usted debe tener en cuenta sobre las librerías:

1. **Conjunto de archivos**. Es uno o varios archivos escritos en un lenguaje de programación determinado, la cual proporciona diversas funcionalidades.
2. **Aporte**. Una librería no aporta la estructura sobre cómo realizar el desarrollo, sino que proporciona funcionalidades comunes, que ya han sido resueltas previamente por otros programadores y evitan la duplicidad de código. Además, reducen el tiempo de desarrollo y aumentan la calidad del mismo.
3. **Espacialidades y fortalezas**. Cada librería se especializa o tiene una fortaleza, por ejemplo, algunas pueden ser muy fuertes en la lectura, manipulación, limpieza y demás acciones con los datos, sin embargo, otro tipo de librerías pueden trabajar de manera más fuerte en las visualizaciones de resultados gráficos.
4. **¿Qué se busca con ellas?** El objetivo es utilizar un conjunto de librerías y que se aprovechen de cada una de ellas sus fortalezas. Dentro del desarrollo de proyectos se puede hacer uso de múltiples librerías y pueden combinarse para obtener mejores resultados.
5. **Las más importantes**. Las librerías más importantes, utilizadas en proyectos de ciencia de datos para la fase de limpieza y transformación de los datos, son:

* Pandas y Numpy: utilizada para la manipulación de grandes volúmenes de datos, cálculos numéricos, análisis de estructuras de datos, análisis estadísticos.
* Matplotlib y Seaborn: utilizada para una mejor comprensión de los datos mediante visualizaciones gráficas y visualizaciones estadísticas.

El video que se muestra enseguida explica cada uno de los pasos para la importación de librerías:

1. Importación de librerías



[Enlace de reproducción del video](https://youtu.be/lSvA7RCXkhM?si=UH4KcPGnc4GpEemm)

|  |
| --- |
| **Síntesis del video: Importación de librerías** |
| En este video, el instructor SENA desarrolla un ejemplo de la importación de las librerías y cómo utilizarlas. Describirá y mostrará los pasos necesarios para este proceso. |

# Preparación de los datos

Una de las etapas más importantes en los proyectos de ciencia de datos es la búsqueda y preparación de los datos, ya que estos son el insumo principal con el que se debe empezar a trabajar. Por ende, es fundamental definir de dónde se obtendrán los datos y qué estrategias se implementarán para la obtención de la información inicial.

Teniendo claros los aspectos anteriores, se puede determinar si los datos privados con los que cuenta la organización son suficientes o si, por el contrario, se necesita obtener información de fuentes externas.

Si el análisis o investigación arroja que se necesitan definitivamente datos externos porque la empresa no cuenta con este insumo, es importante iniciar dicha preparación planteándose algunas preguntas, como:

* ¿Se pueden conseguir, los datos externos, en fuentes públicas?
* ¿Estos datos públicos son gratis o tienen algún costo?
* ¿Qué tan fiables son las fuentes de información de donde obtendría estos datos?
* ¿En qué tipo de formato puedo obtener la información?

Una vez resueltas todas las dudas que se puedan presentar en la obtención de los datos, es importante determinar la estrategia para recuperar esa información de manera eficiente, dicha estrategia debe contemplar, los recursos tecnológicos y humanos que se requieren para satisfacer ese requerimiento.

Por ejemplo, en este punto, es un deber que dentro de la organización se cuente con personal capacitado en:

* El manejo de bases de datos relacionales y no relacionales.
* Manejar el lenguaje SQL para la extracción personalizada de información.
* Manejar los diferentes formatos y estructuras para almacenar la información.

Algunas fuentes de información públicas requieren de registro en sus plataformas para descargar información, otras simplemente requieren de ciertos datos para descargarla sin ninguna restricción.

Ejemplo búsqueda de información

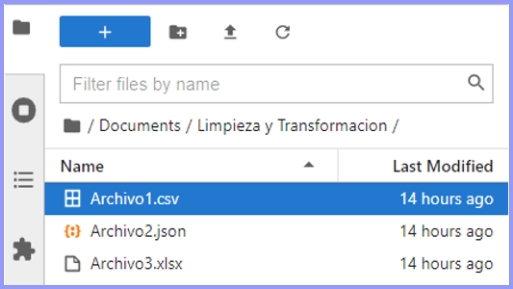
Consulte el PDF denominado **Ejemplo\_busqueda\_de\_informacion,** allí encontrará un ejemplo de búsqueda de información en fuentes de datos que son públicas y que se pueden descargar gratuitamente. Dicho archivo se encuentra en la carpeta Anexos.

### Lectura de archivos en diferentes formatos

A continuación, a través de un ejemplo, se van a leer diversos tipos de formatos y estructuras, se revisarán las estructuras de cada uno de los archivos y se procederá con la unificación, entendiendo que todos los archivos manejan el mismo número de columnas en la misma posición.

Para comenzar, es importante ubicar, en la misma ruta donde se crearán los “notebooks”, tres tipos de archivos diferentes como se muestra en la figura siguiente:

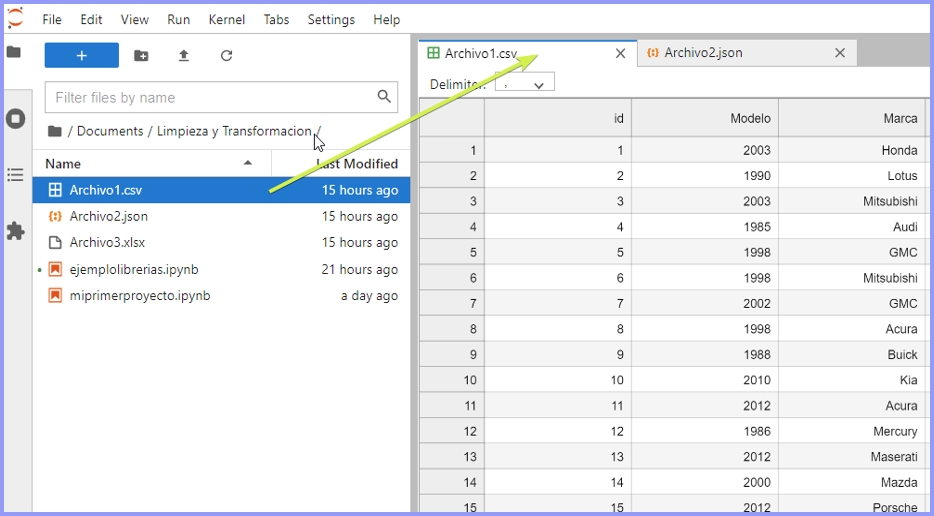
1. Carpeta con archivos



Como la imagen muestra, se trata de archivos diferentes, identificados cada uno con su respectiva extensión, el primer archivo llamado Archivo1.csv es un archivo delimitado por comas con extensión CSV, segundo archivo Archivo.json tal como su extensión lo indica es un archivo con una estructura tipo JSON, y el último archivo denominado Archivo.xlsx es una hoja de cálculo con formato XLSX y como se sabe es una extensión de archivos del programa Excel.

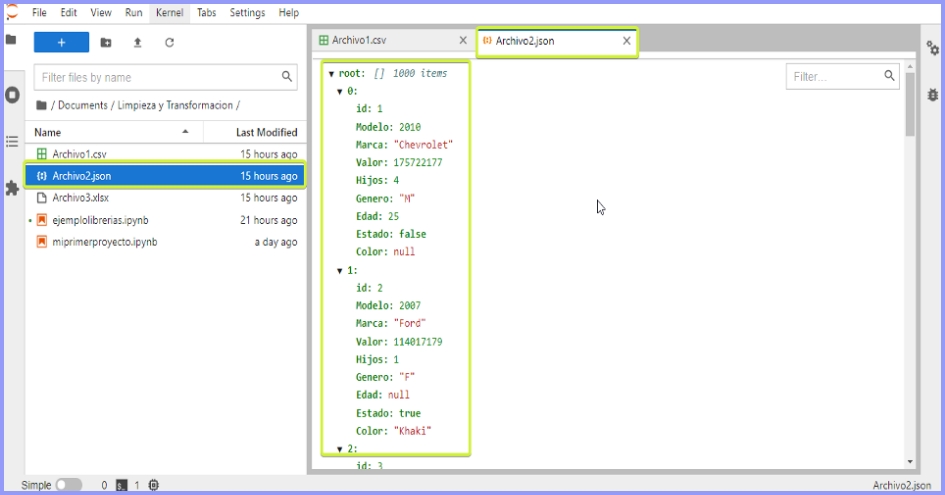
Directamente desde la plataforma JupyterLab se pueden ver los archivos para analizar un poco los datos, para ese fin es importante hacer clic sobre el archivo que desea visualizar y la información se mostrará en el panel de la derecha:

1. Visualizando archivos desde JupyterLab



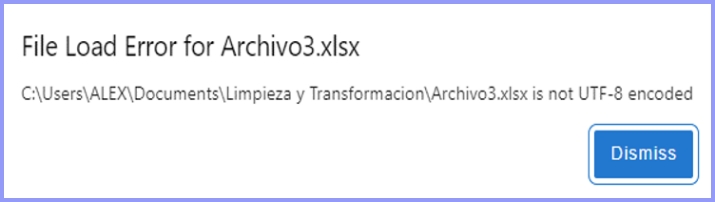
Para ver el archivo tipo JSON, igualmente se debe hacer doble clic sobre el archivo con extensión JSON:

1. Visualizar archivos tipo JSON



Al leer el último archivo que es de tipo XLSX puede presentar el siguiente error:

1. Error abriendo el archivo de Excel



Este archivo no se puede abrir directamente en el panel, aunque esto no sería una preocupación, toda vez que este tipo de archivo se puede visualizar directamente con el programa Excel, pero sí se debe tener en cuenta la codificación UTF-8 al momento de leerlo para pasarlo a un “Dataframe”.

Para proceder con la lectura de cada uno de los archivos y unificarlos, para preparar un “Dataframe” y posterior limpieza y transformación de los datos, se deben realizar unos pasos específicos, como los que se detallan enseguida:

1. **Crear un nuevo “notebook”**. Crear una nueva área de trabajo donde se empezará con la importación de las librerías necesarias para la lectura de los archivos. Es necesario ingresar las siguientes librerías:

* import pandas as pd
* import numpy as np
* import matplotlib.pyplot as plt
* import seaborn as sns
* from io import open

1. **Leer archivo CSV**. Para realizar la lectura de archivos con extensión CSV, se ejecuta la siguiente línea de comandos, el resultado muestra los cinco primeros y cinco últimos registros importados, al final se debe visualizar el total de registros y número de columnas:

* df1 = pd.read\_csv('Archivo1.csv')
* df1

1. **Importar archivos JSON**. Este tipo de estructuras se deben convertir en filas y columnas para que se puedan utilizar en el análisis de los datos y poderlas unir, también, con los otros dos archivos. Para realizar la importación de este tipo de estructuras se debe usar el siguiente comando:

* df2 = pd.read\_json('Archivo2.json')
* df2

1. **Importar archivos XLSX**. Para realizar la importación de este archivo se debe tener cuenta el error que se generó anteriormente. Para solucionar este problema y poderlo importar sin inconvenientes, se deben ejecutar los comandos de la siguiente manera: primero se realiza la codificación para luego realizar la importación, de lo contrario se generarán errores.

* f = open("Archivo3.xlsx", mode="r", encoding="utf-8")
* df3 = pd.read\_excel('Archivo3.xlsx')

1. **Unir los “dataframe”**. Cada archivo se importó en un “Dataframe” independiente, para este ejemplo, cada uno de los resultados de las importaciones manejan las mismas columnas en las mismas posiciones, lo que quiere decir que se puede unificar en un solo archivo; cada archivo cuenta con mil registros y nueve columnas, por ende, el resultado final debe mostrar 3000 registros y nueve columnas, para obtener este resultado se debe ejecutar las siguientes líneas de comando:

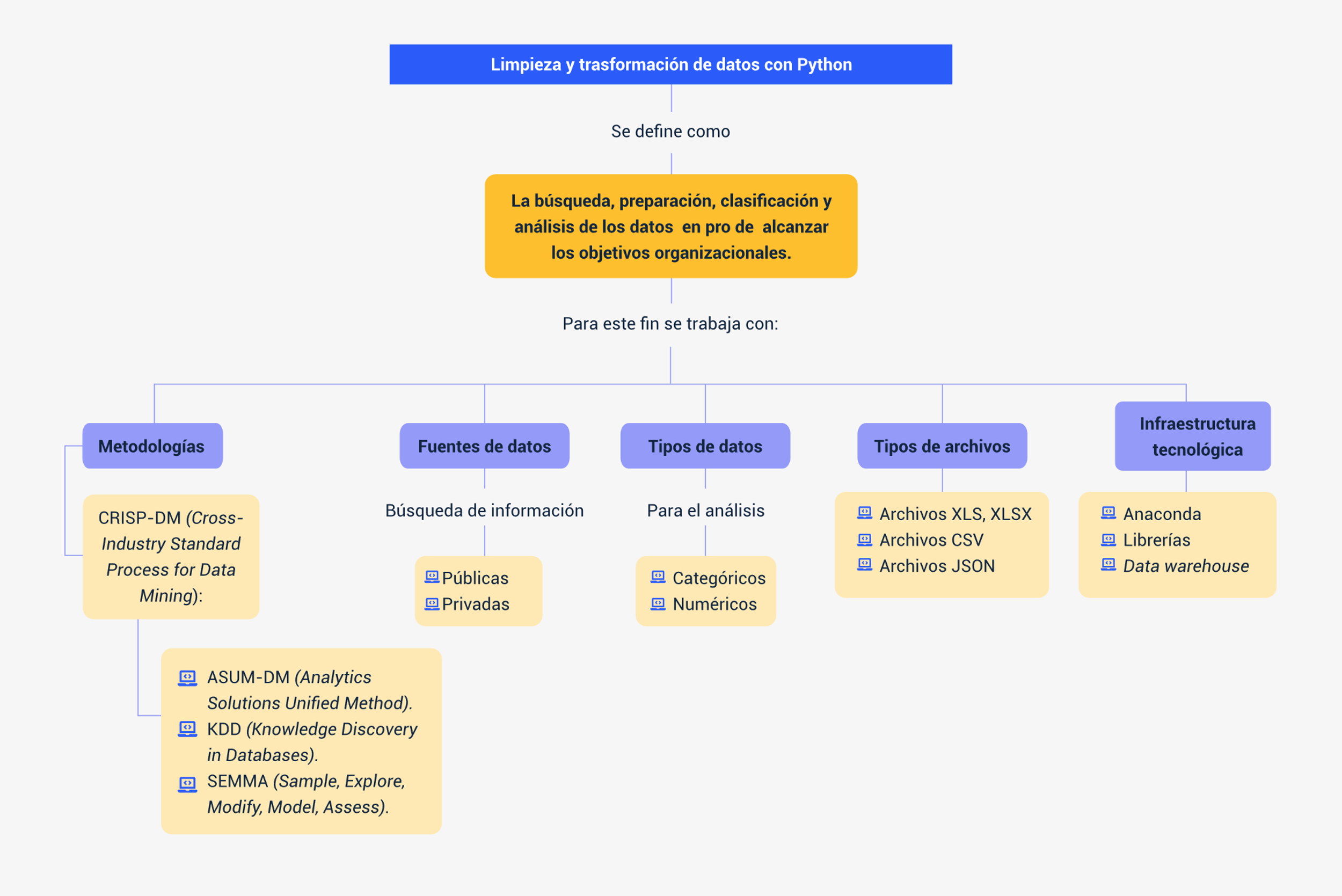
* df=pd.concat([df1,df2,df3])
* df

1. **Guardar en único formato**. Se debe guardar todo el proceso en un único tipo de formato para que en las próximas lecturas solo se tenga que leer un solo archivo con toda la información consolidada, para ello es importante ejecutar la siguiente línea de comando:

* df.to\_csv('Unificado.csv')

Síntesis

Aquí finaliza el estudio de este componente formativo. En este punto, preste atención al esquema que se muestra enseguida y realice su propia síntesis de los temas estudiados. ¡**Adelante**!



Este es el mapa de conceptos y temáticas desarrollados en este componente formativo. Su tema principal: la limpieza y transformación de datos con Python. Sus temas integradores: **metodologías de análisis de los datos**, **fuentes de datos** (que vincula tanto tipos de bases de datos como tipos de archivos), **infraestructura tecnológica** (que abarca la instalación, la puesta en marcha y la instalación e importación de librerías), **preparación de datos**.

Material complementario

|  |  |  |  |
| --- | --- | --- | --- |
| Tema | Referencia | Tipo de material | Enlace del recurso |
| 1. Metodologías de análisis de datos. | Conceptos básicos de ayuda de CRISP-DM. (2022). Metodología Crisp-DM. | Página Web | <https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview> |
| 2.1 Tipos de base de datos. | Holamyndo. (s.f). ¿Qué es SQL y NOSQL? cuáles son sus diferencias y cuándo deberías utilizarlos. [Video]. YouTube | Video | <https://www.youtube.com/watch?v=zmXl2dOGWL8&t=9s> |
| 3. Infraestructura tecnológica | Pandas documentation — Pandas 1.4.2 documentation. (2022). Pandas. | Página Web | <https://pandas.pydata.org/docs/index.html> |

Glosario

**Base de datos**: sistema que permite el almacenamiento de datos de manera organizada ya sea de manera relacional o no relacional.

**CSV**: extensión para definir archivos delimitados por comas, no almacena texto enriquecido ni imágenes o tablas.

**Extensión**: identificador al final de cada archivo que define su estructura.

**JSON**: formato que se basa en una estructura de texto simple que almacena la información en forma de objetos.

**Lenguaje SQL**: se utiliza para recuperar, modificar o eliminar información de una base de datos relacional.

**Librerías**: conjunto de instrucciones que realizan una labor específica, y facilitan el desarrollo de proyectos.

**Tipo de archivo**: dependiendo del tipo de archivo se puede almacenar la información en estructuras diferentes de acuerdo con la necesidad o “software” que lo cree.

**XLSX**: extensión que se utiliza para identificar archivos procedentes de hojas de cálculo, como por ejemplo Excel.

Referencias bibliográficas

Datos Abiertos Colombia | Datos Abiertos Colombia. (2022). La plataforma de datos abiertos del gobierno colombiano. <https://www.datos.gov.co/>

Introducción al CRISP-DM. (2021). IBM. <https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=guide-introduction-crisp-dm>

L. (2021). Manual de Anaconda y Conda. ¿Cómo administro y configuro mi entorno para Python? - Tutoriales de programación y un poco más. <https://tutorialdeprogramacion.com/manual-conda-y-python/>

NumPy. (2022). Numpy. <https://numpy.org/>

Oracle. (2022). Base de Datos. <https://www.oracle.com/co/database/what-is-database/#:%7E:text=Una%20base%20de%20datos%20es,bases%20de%20datos%20(DBMS)>

Oracle. (2022). Tipos de Bases de Datos. <https://www.oracle.com/co/database/what-is-database/#link5>

Tiempo, A. (2022,). Limpieza de datos con Python - Al mal tiempo, buena data. Medium. <https://lauralpezb.medium.com/limpieza-de-datos-con-python-48d436ca9ace>

Créditos

| Nombre | Cargo | Centro de Formación y Regional |
| --- | --- | --- |
| Claudia Patricia Aristizábal | Responsable del Ecosistema | Dirección General |
| Rafael Neftalí Lizcano Reyes | Responsable de Línea de Producción | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Ronald Alexander Vacca Ascanio | Experto Temático | Dirección General, Centro de Diseño y Metrología |
| Jeimy Lorena Romero Perilla | Diseñador Instruccional | Regional Norte de Santander - Centro de la industria, la empresa y los servicios |
| Carolina Coca Salazar | Asesora Metodológica | Distrito Capital- Centro de Diseño y Metrología |
| Jhon Jairo Rodríguez Pérez | Corrector de estilo | Regional Distrito Capital - Centro de Diseño y Metrología |
| Fabián Leonardo Correa Díaz | Diseñador Instruccional | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Blanca Flor Tinoco | Diseñador de Contenidos Digitales | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Emilsen Alfondo Bautista | Desarrollador “Fullstack” | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Wilson Andrés Arenales Cáceres | Storyboard e Ilustración | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Carlos Eduardo Garavito Parada | Animador y Productor Multimedia | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Camilo Andrés Bolaño Rey | Locución | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Zuleidy María Ruiz Torres | Validador de Recursos Educativos Digitales | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Luis Gabriel Urueta Álvarez | Validador de Recursos Educativos Digitales | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Daniel Ricardo Mutis Gómez | Evaluador para contenidos inclusivos y accesibles | Regional Santander - Centro Industrial del Diseño y la Manufactura |