



AA SELECCIONA LA MUESTRA PARA ANALITICA DE DATOS (TI)

IDENTIFICACIÓN DE LA GUÍA DE APRENDIZAJE

- Denominación del Programa de Formación: TÉCNICO PARA LA PROGRAMACIÓN PARA ANALÍTICA DE DATOS.
- Código del Programa de Formación: 228117
- Nombre del Proyecto: 1901119 APLICAR BUENAS PRACTICAS PARA PREPARAR, LIMPIAR, REFINAR Y EXPLORAR GRANDES VOLÚMENES DE DATOS EN EL SECTOR PRODUCTIVO.
- Fase del Proyecto: PREPARACIÓN DE DATOS, EXPLORACIÓN DE DATOS, ADQUISICIÓN DE DATOS
- Actividad de Proyecto:
 - ORGANIZAR LA DATA
 - GENERAR INFORME
 - SELECCIONAR DATOS
- Competencia:
 - 220501115 INTEGRAR DATOS SEGÚN TÉCNICAS DE VISUALIZACIÓN Y METODOLOGÍAS DE ANÁLISIS.
- Resultados de Aprendizaje Alcanzar:
 - RAP 45 ORGANIZAR LA INFORMACIÓN A GESTIONAR DE ACUERDO CON TÉCNICAS DE ANÁLISIS.
 - RAP 46 ELABORAR INFORMES UTILIZANDO HERRAMIENTA INFORMÁTICA SELECCIONADA.
 - RAP 50 RECOLECTAR INFORMACIÓN DE ACUERDO A LAS NECESIDADES DEL CLIENTE.
 - RAP 49 ELABORAR INFORMES SEGÚN LA NECESIDAD DEL CLIENTE.
- Duración de la Guía: 168 (presenciales + trabajo autónomo)

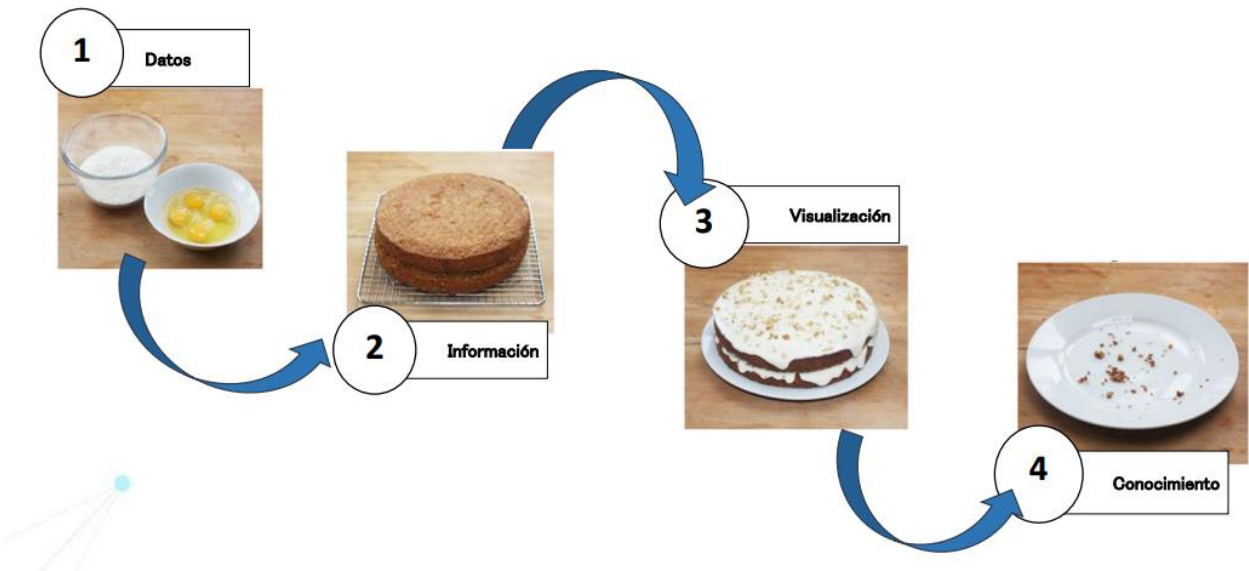
2. PRESENTACIÓN

Para los proyectos de analítica de datos es importante conocer las diferentes técnicas, arquitecturas y fuente de datos, en esta actividad de aprendizaje abordaremos estos temas para que se adquieran las habilidades necesarias para participar en proyectos de analítica e inteligencia de negocios.

3. FORMULACIÓN DE LAS ACTIVIDADES DE APRENDIZAJE

3.1 Actividades de Reflexión inicial.

Para iniciar el proceso reflexionaremos sobre los términos de datos e información, centrado en el proceso que se debe realizar para convertir el dato en información, teniendo en cuenta los diferentes tipos de fuentes y formatos; para esto se dispone de un foro donde se plasme de manera sucinta sus opiniones al respecto, no olvide debatir sobre los comentarios de sus compañeros.



Con la anterior ilustración, se debe dar respuesta a las siguientes preguntas:

- ¿Qué es dato?
- ¿Qué es información?
- ¿Cuál es la diferencia entre dato e información?
- ¿Cuáles fuentes de información conoce?

De su opinión en el foro dispuesto en la plataforma LMS sobre las preguntas planteadas.

Embudo del Conocimiento



Fuente: Jose Aguilar, CEMISID, aguilar@ula.ve

“Todo debe hacerse lo más simple posible. Pero no más sencillo.”
Albert Einstein

3.2 Actividades de contextualización e identificación de conocimientos necesarios para el aprendizaje.

Para un proyecto de analítica de datos debemos tener en cuenta las siguientes fases:

Localizar: Donde se encuentra la información

Adquirir: Como se va a acceder a la información: por archivos planos, driver de conectividad (odbc, jdbc, nativos), hojas electrónicas, datos internos, datos externos, etc.



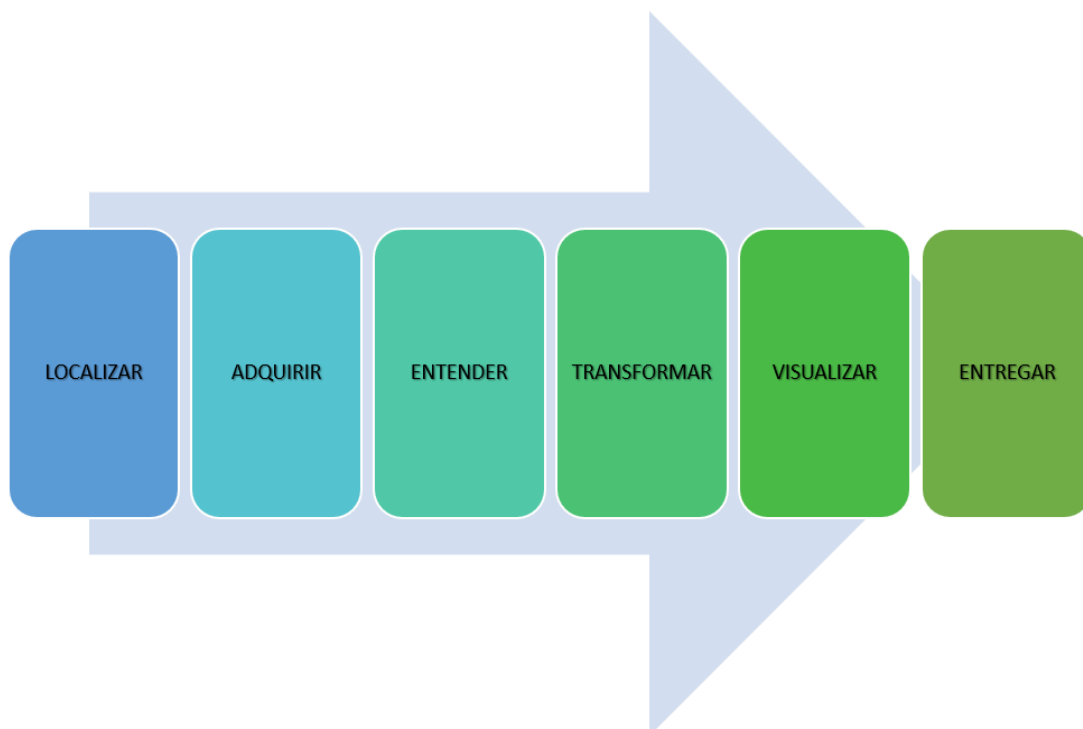
Entender: De acuerdo a las pretensiones del proyecto se debe determinar las dimensiones que se va a procesar, teniendo en cuenta el Habeas Data; el diccionario de datos para entender los dominios de los datos a procesar.

Transformar: Algunos datos vienen con formato de fechas no compatibles, datos perdidos, datos no categorizados, errores de ortografía, caracteres escondidos, etc.; Para esto se recomienda la guía QUARTZ de limpieza de datos.

Visualizar: Una vez que se ha procesado la data y exportada la muestra para ser entregada es necesaria visualizar de acuerdo a las técnicas estadísticas para determinar que los datos están correctos y categorizados.

Entregar: Se debe entregar la muestra de acuerdo a un acta de entrega, teniendo en cuenta el diccionario de datos, el medio, las transformaciones realizadas y las gráficas de visualización.

¿Por qué es importante saber identificar fuentes de información, la manera como se adquiere, entender los datos adquiridos, procesarlos y visualizarlos para proyectos de analítica de datos?

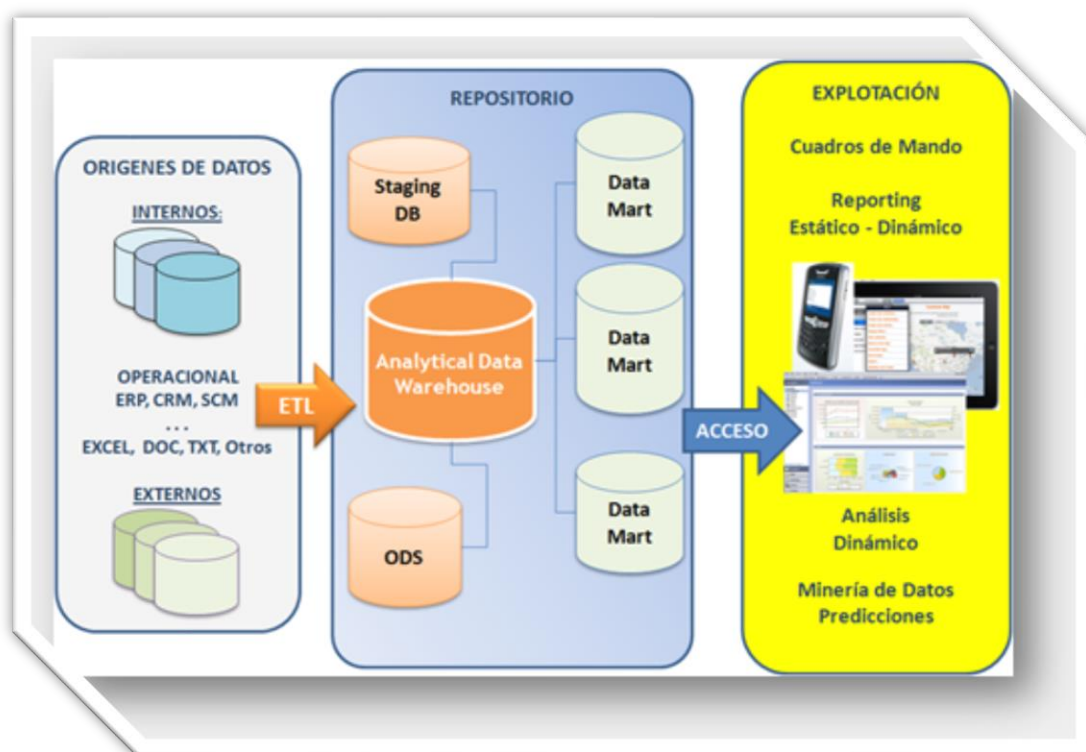




Se debe garantizar la seguridad de la información y su resguardo efectivo para cuando ocurra un suceso inesperado o corregir vulnerabilidades y riesgo a éstos.

Nunca se puede resolver un problema en el mismo nivel en el que fue creado.
Albert Einstein

3.3 Actividades de apropiación del conocimiento (Conceptualización y Teorización).



Jurisprudencia de analítica de datos

- [Compes 3920](#)
- [Habeas Data](#)
- [Datos abiertos](#)
- [Datos abiertos ICFES](#)

Fuentes de datos

- Internas
- Externas
- [Datos abiertos](#)

[Clase de fuentes](#)



- Primarias
- Secundarias
- Terciarias
- Estadísticos



Tipos de datos

- Estructuradas
- Semi estructuradas
- No estructurados

Recolección de datos

- Batch o por lotes
- Streaming o transmisión en tiempo real
- Técnicas de recolección de datos

Formatos de los datos

- Planos
 - Delimitados (CSV)
 - Ancho fijo
 - JSON, BJSON
 - XML
- Base de datos
- Hojas electrónicas



- [PDF](#)
- Imágenes, videos y audios.
- Logs de un servidor
- Métricas del sistema
- Tweets de un hashtag.
- Sentimiento general en comentarios de redes sociales.
- Histórico de visitas de clientes a una tienda.



Para iniciar a trabajar con [Python](#) se recomienda ver el video que le ayudara a contextualizar sobre el lenguaje de programación.

En compañía del instructor se recomienda realizar el taller [FUNDAMENTOS DE PROGRAMACIÓN EN PYTHON](#).



No olvide de guardar una copia en su portafolio de evidencias.



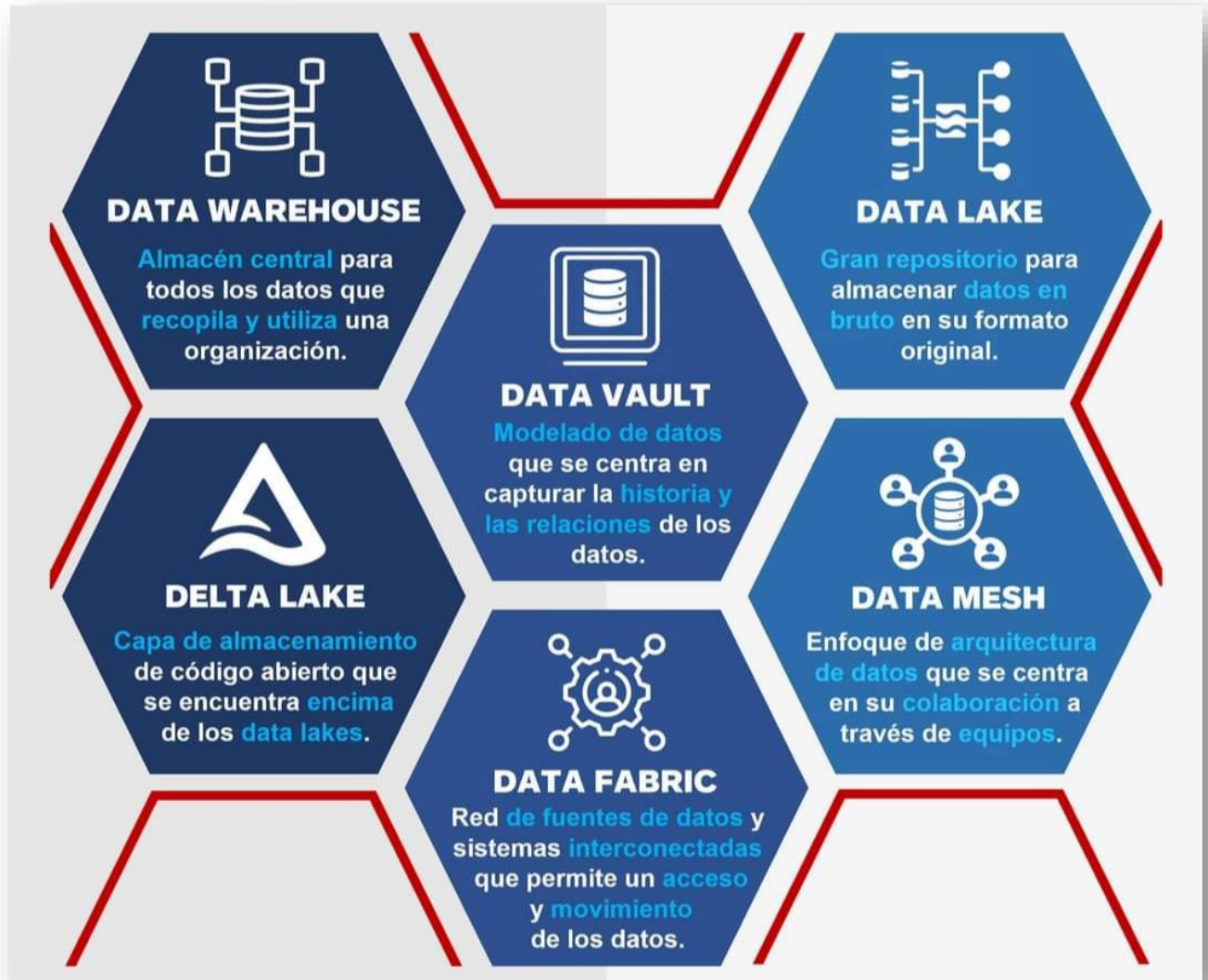
Realizar el taller de [Neptuno](#), utilizando los [datos](#) que se adjuntan.



No olvide guardar una copia en su portafolio de evidencias y subir la evidencia que consiste en convertir cada hoja del libro [Neptuno](#), comprimir y subir al LMS.

Se recomienda ver el video: [¿Cómo cargar y descargar archivos CSV desde Google Colab?](#)

El instructor expone un ejemplo de [dataframe](#) en Python



En compañía del instructor se recomienda realizar el taller MODELOS DE INTELIGENCIA DE NEGOCIO

¿Qué es BIGDATA?



Se recomienda ver el video: “[¿Qué es BIG DATA y para qué sirve?](#)” y “[Curso Analista de datos](#)”

¿Qué es BIGDATA?

Se recomienda las siguientes lecturas que se encuentran en los materiales de formación:

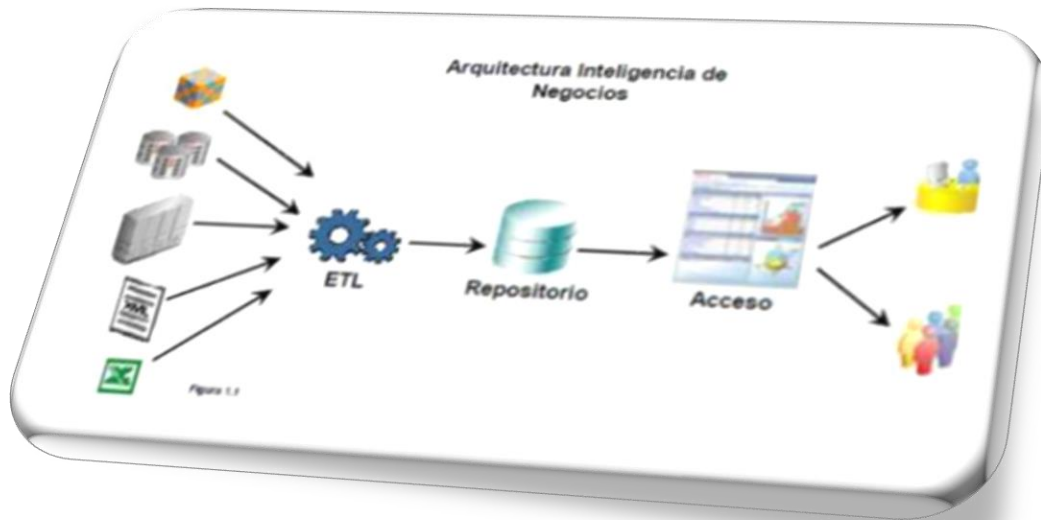
- Revisar la infografía “[RECONOCER LOS COMPONENTES DE BIG DATA\(HADOOP\)](#)”.
- INTRODUCCIÓN AL BIGDATA, INTELIGENCIA DE NEGOCIO Y FUENTES DE DATOS
- OPEN DATA, Arquitectura BIG DATA
- INTRODUCCIÓN A LA MINERÍA DE DATOS
- PREPARACIÓN DE DATOS
- MINERÍA DE DATOS – PREPROCESAMIENTO
- ANÁLISIS EXPLORATORIO
- METODOLOGÍAS
- PROCESOS ETL

Proceso ETCL

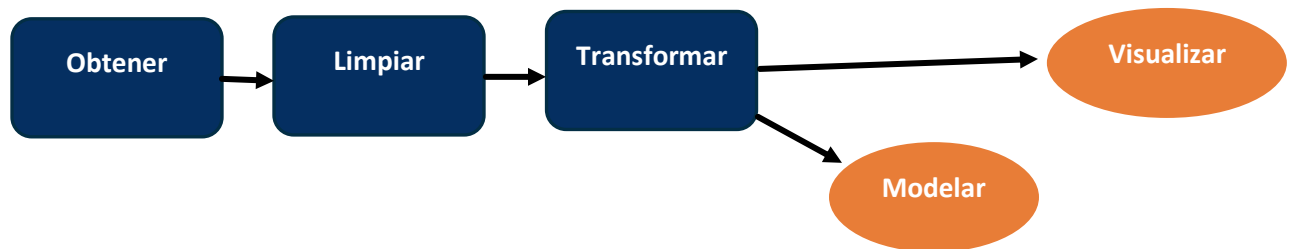
- Extracción (E)
- Transformación (T)
- Limpieza (C)
- Carga (L)



Se recomienda hacer la lectura “[Procesos ETL de PowerData](#)”



Se denomina “data wrangling o data munging” al proceso de transformación significa “peleando o riñendo con los datos” y el segundo a cambios irreversibles que se realizan a un conjunto de datos para transformarlos en otros distintos.



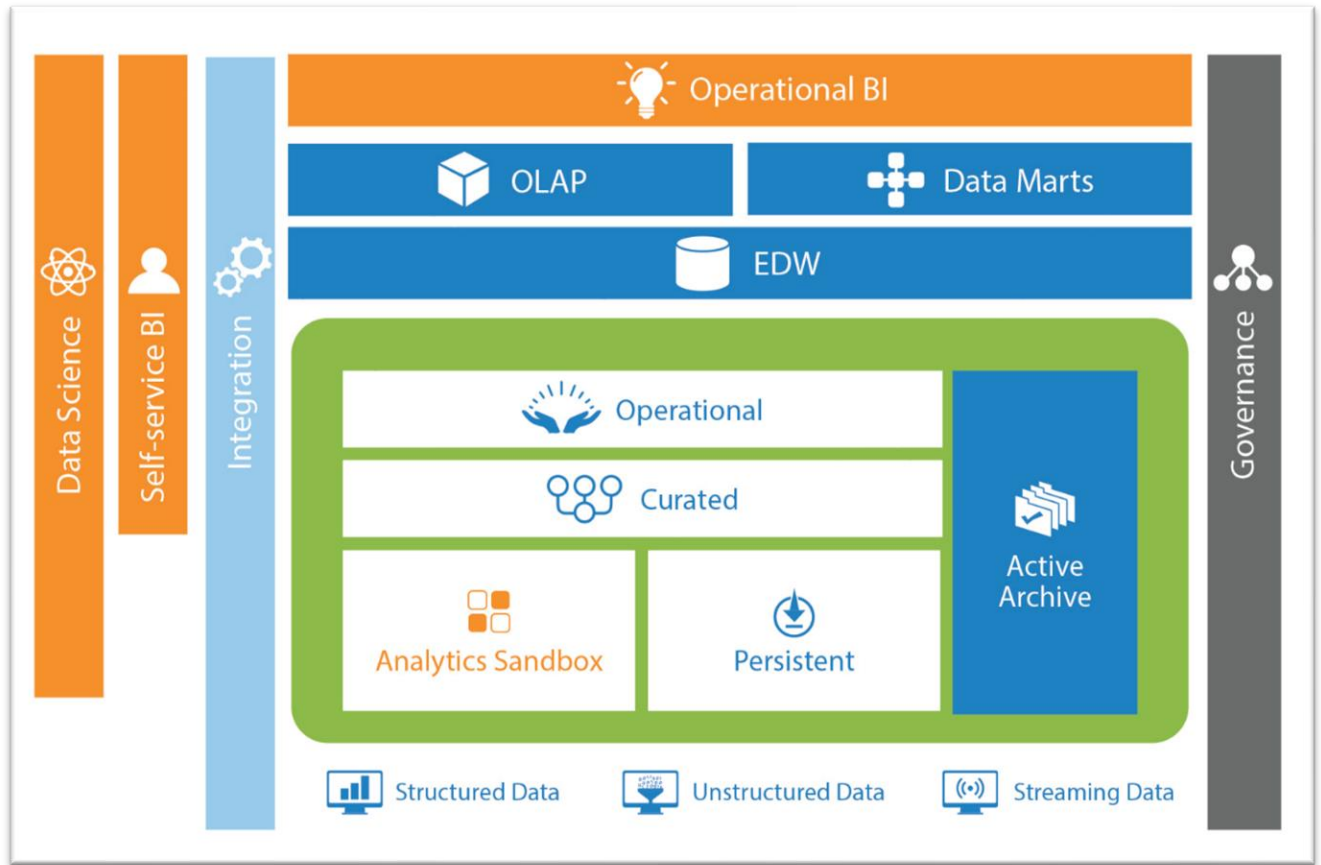
El data wrangling es muy similar al proceso ETL, en las transformaciones, pero diferente en algunos procesos como es adecuar los datos en crudo aplicando técnicas de extracción de información que se aplicaran sobre ellos como son:

- Muestrear datos
- Crear nuevos atributos
- Discretizar atributos cuantitativos continuos
- Fusionar y reordenar datos y atributos
- Completar datos
- Re escalar magnitudes



- Creación de variables dummy (variables indicadoras)
- Reducción de la dimensionalidad
- Eliminación de datos espurios.

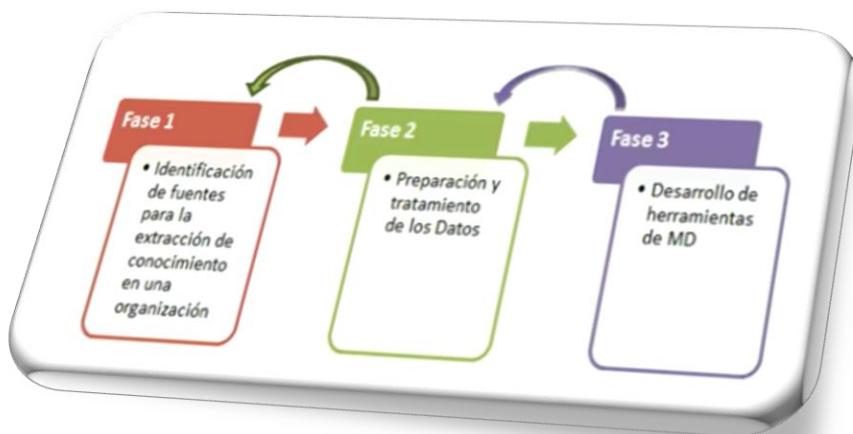
Arquitectura de analítica de datos e inteligencia de negocios



- Capa de fuentes de datos
- Capa de proceso ETL
- Capa de almacenes de datos (DataWareHouse, DataMart)
- Capa de metadatos
- Capa de usuario final (Análisis y visualización de resultados)

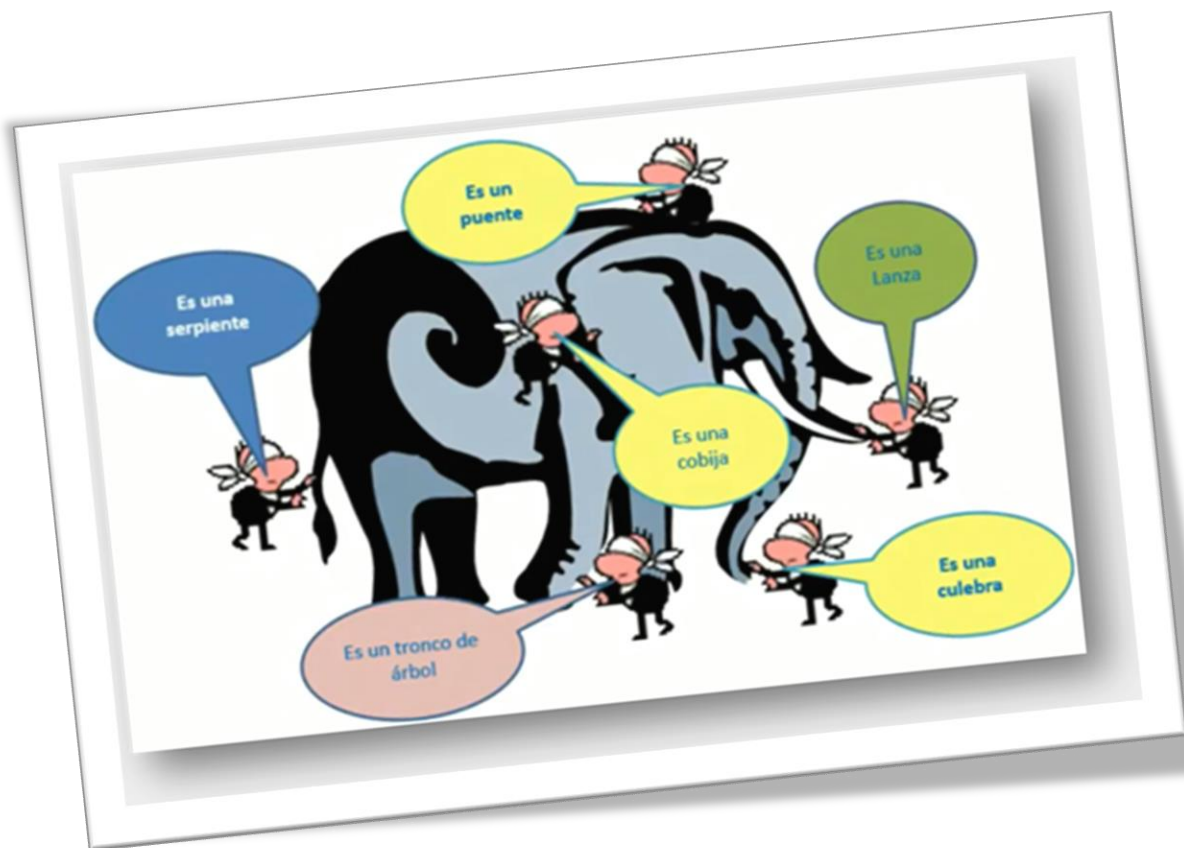
Se recomienda ver el video "[Arquitectura de una bodega de datos](#)"

El modelo MIDANO tiene como énfasis determinar la especificación de tareas de analítica de datos



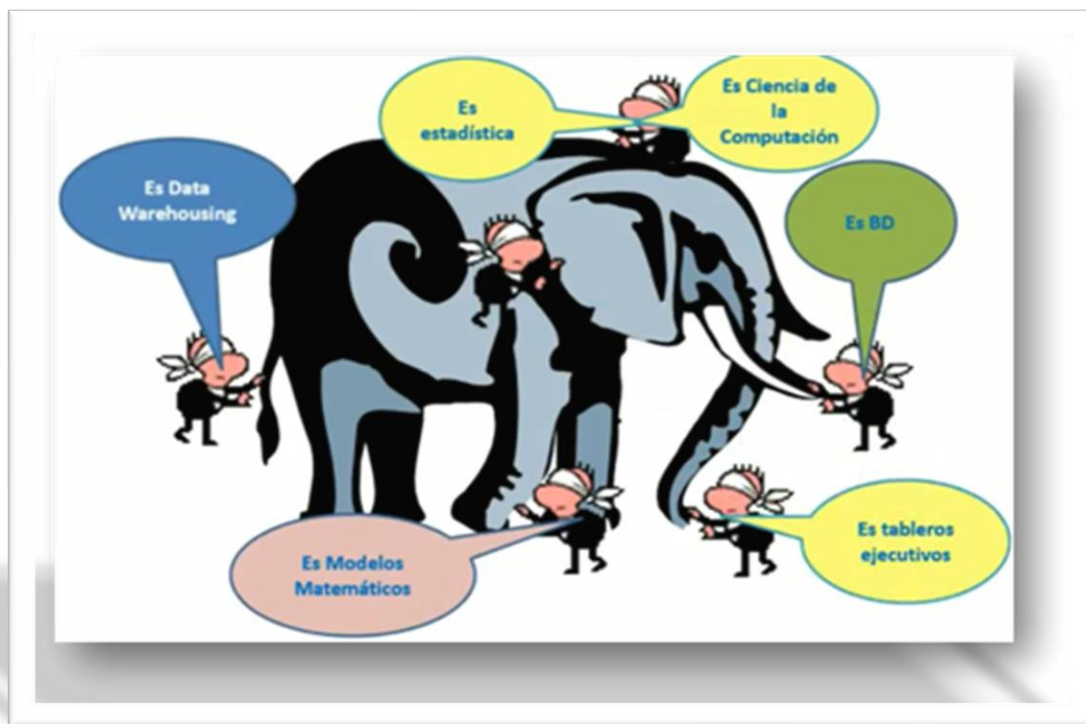
Fuente: Jose Aguilar, CEMISID, aguilar@ula.ve

Un ciego describiendo un elefante.





¿Cómo describimos a BIG DATA?



Fuente: Jose Aguilar, CEMISID, aguilar@ula.ve

Limpieza de datos es el proceso en los cuales la calidad de los datos es mejorada, como: datos mal capturados, anómalos, vacíos, formatos, caracteres escondidos, ortografía, duplicados, abreviaciones, unidades de entradas y otros.

Transformación proceso por el cual las variables de entrada se transforman en nuevas variables de interés, por ejemplo: concatenación de cadenas, multiplicación entre variables, desagregación de cadenas.

Reducción decidir qué datos deben ser utilizados para el análisis, utilizando análisis estadístico, identificar las posibles variables que se pueden reducir, justificar la reducción de las mismas y construir la nueva data con las variables definitivas.

OLAP (Procesamiento Analítico en Línea)

Se realiza con herramientas de análisis multidimensionales. El creador de la idea OLAP, minería de datos y sistema de apoyo a la decisión (DSS) fue de Edgar Frank Codd.

Algunas características de los sistemas OLAP son:



- Permite visualizar los mismos datos en diferentes sitios, utilizando múltiples dimensiones.
- Respuestas rápidas a problemas ad hoc.
- Alto nivel de detalles en cada operación.

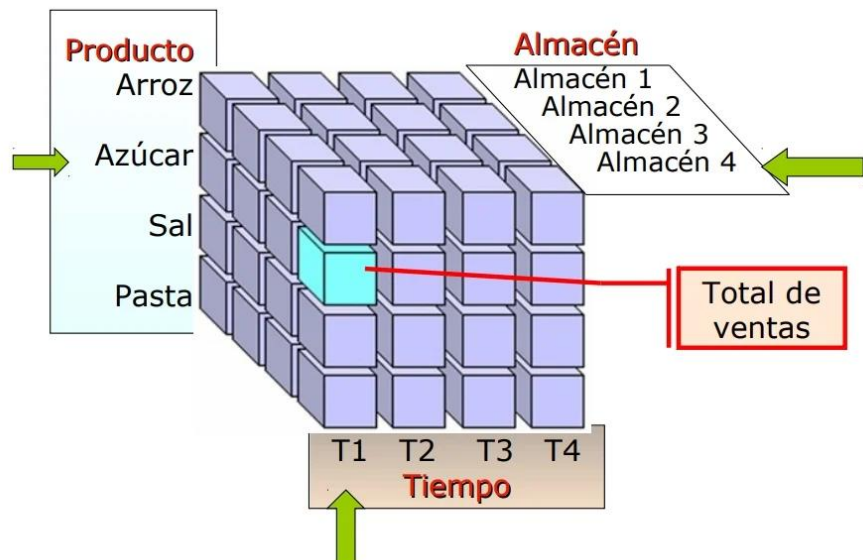
Los elementos de control son:

- Variables de decisión: representan una medición del negocio.
- Basados en el concepto de cubos de datos OLAP
- Las herramientas OLAP se caracterizan por:
 - Ofrecen una visión multidimensional
 - No hay restricciones sobre el número de dimensiones.
 - Ofrece simetría para las dimensiones.
 - Permite definir de forma flexible las dimensiones, restricciones, agregaciones y jerarquías entre ellas.
 - Ofrecen operadores intuitivos de manipulación (drill-down, roll-up, slice-and-device, pivot, etc.)
 - Son transparentes en otras tecnologías como ROLAP y MOLAP

CUBOS OLAP

Es una base de datos que posee varias dimensiones, que amplía las posibilidades que se hacían con las hojas electrónicas.

Cada una de las dimensiones incorporan un campo determinado para un tipo de dato específico, que se podrá comparar con la información contenida en el resto de las dimensiones para evaluar la información relevante.



MOLAP

Se implementa en una base de datos multidimensional, los datos se organizan en una estructura tipo cubo que el usuario puede rotar. Utilizadas para resúmenes e informes financieros.

ROLAP

Se implementa en una base de datos multidimensional, accediendo directamente a los datos almacenados en un Data Warehouse para proporcionar los análisis especificados. Puede crear vistas multidimensionales, pero no en estructuras tipo cubo.



HOLAP (OLAP HIBRIDA)

Almacena datos en un sistema de base de datos relacional y otros en base de datos dimensionales, tratando de combinar las ventajas de MOLAP y ROLAP .



Fuente: <https://www.northware.mx/blog/que-es-un-modelo-de-analitica-de-datos/>

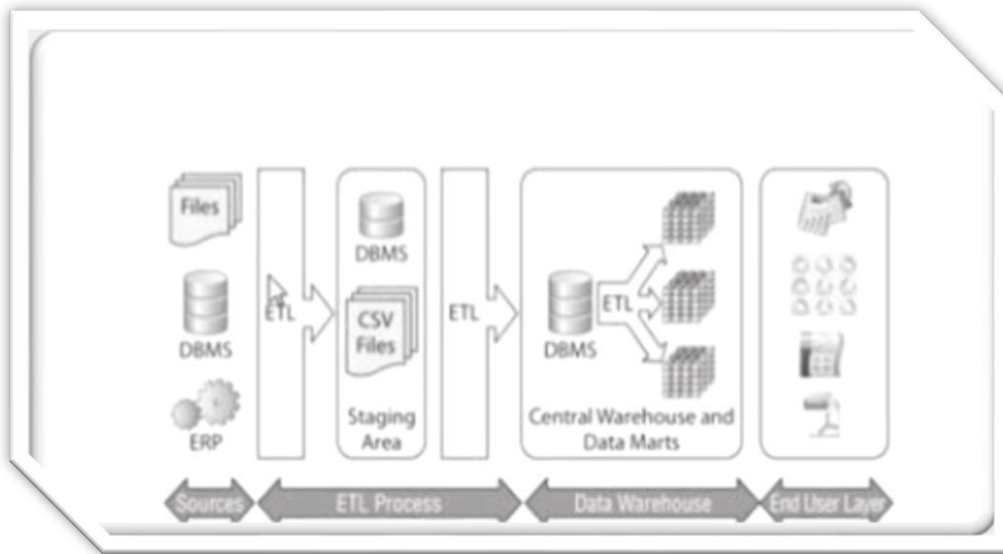
Proceso de analítica de datos:

- Planteamiento de la hipótesis inicial, ¿cómo hacer la hipótesis?
- Preparación y procesamiento de datos
- Análisis exploratorio de datos
- Transformación de variables
- Reducción de dimensiones (Feature Engineering)

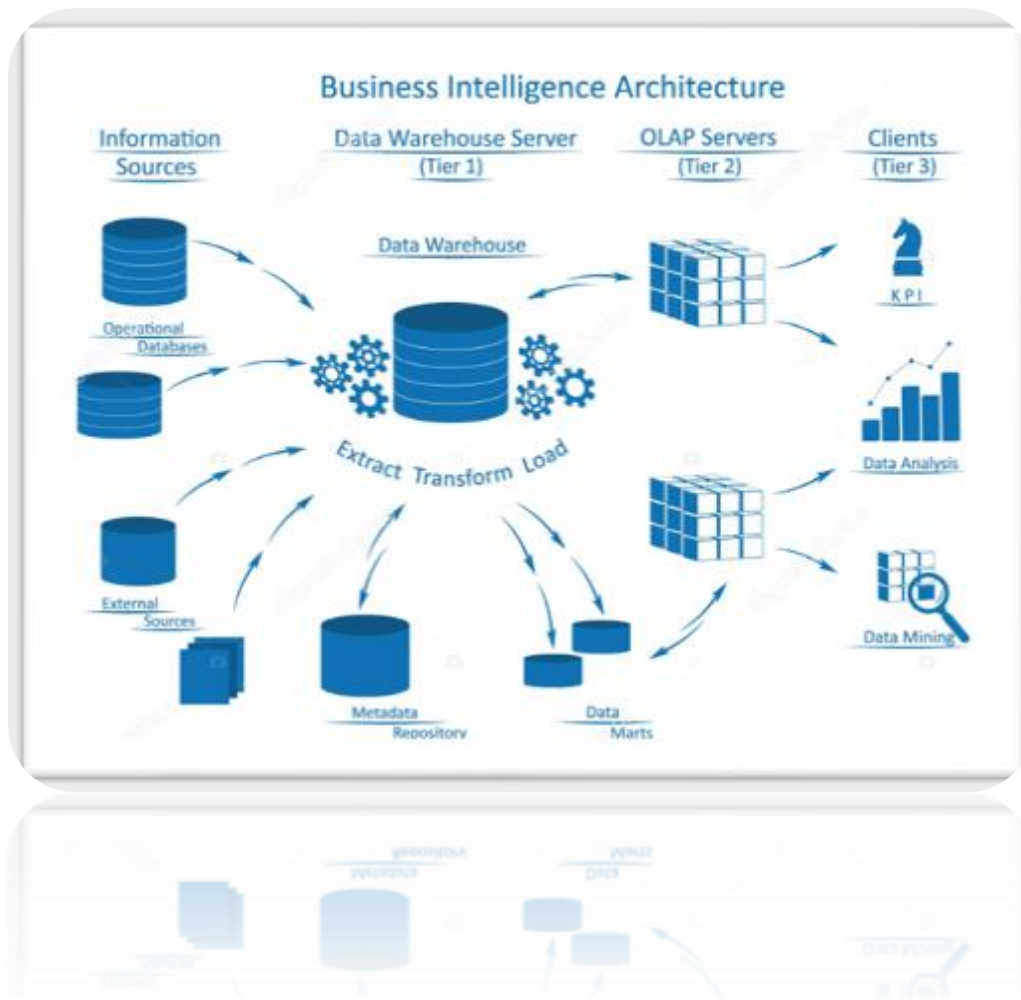
Tipos de base de datos

- Relacionales
- No relacionales (NOSQL)
 - Orientadas a columnas
 - Orientadas a documentos
 - Orientadas por clave – valor
 - Orientadas a grafos

Se recomienda ver la siguiente infografía de [base de datos no relacionales](#).



Se recomienda ver el video "[Diferencia entre inteligencia de negocios y analítica de datos](#)", "[Anaítica de datos, Ciencia de datos, Machine Learning y Inteligencia Artificial](#)"





Modelo de datos utilizadas en inteligencia de negocios

- [Modelo relacional](#)
- [Modelo estrella](#)
- [Modelo copo de nieve](#)

“Una vez que aceptamos nuestros límites, podemos ir más allá de ellos.”
Albert Einstein

Herramientas informáticas para analítica de datos

- [Python](#) 
para analítica de datos
- [R](#) y [RSTUDIO](#) 
para analítica de datos. Se puede ejecutar en [línea](#)



- Fundamentos de [WEKA](#) ↓
para analítica de datos
- Fundamentos de [ORANGE](#) ↓
para analítica de datos
- ETL [RapidMiner](#) ↓
- Fundamentos de [PSPS](#) ↓
para analítica de datos
- [OpenRefine](#) ↓
para limpieza de datos

Se recomienda ver los siguientes tutoriales:

- [Python](#)
- [R](#)
- [WEKA](#)
- [ORANGE](#)
- [RAPIDMINER](#)
- [PSPS](#)
- [OPENREFINE](#)

“La debilidad de actitud se vuelve debilidad de carácter.”
Albert Einstein

3.4 Actividades de transferencia del conocimiento

Para comprobar el aprendizaje realizaremos un caso de estudio que nos permitirá demostrarlo:

se hace necesario realizar lo siguiente:

- Construir un diseño de base de datos del [TIPO ESTRELLA](#) que permita migrar los datos de la fuente de datos del [COVID19](#), mediante una tabla de hecho y sus dimensiones.
- Realizar el modelo físico de acuerdo con el diseño de base datos de la fuente de datos.



“Hay una fuerza motriz más poderosa que el vapor, la electricidad y la energía atómica: la voluntad.”

Para desarrollar la actividad del caso de estudio propuesto se debe realizar el taller [“CASO DE ESTUDIO COVID19”](#) y acompañado con el instructor que le guíara con los conocimientos necesarios para lograr construir la evidencia.



Esta evidencia se debe elaborar en grupo de tres aprendices en las fechas acordadas en el plan de trabajo concertado (PTC) y las instrucciones dadas por el instructor.



No olvide guardar la evidencia en el portafolio del aprendiz.

“Nunca consideres el estudio como una obligación sino como una oportunidad para penetrar en el bello y maravilloso mundo del saber”
Albert Einstein

4. ACTIVIDADES DE EVALUACIÓN

Evidencias de Aprendizaje	Criterios de Evaluación	Técnicas e Instrumentos de Evaluación
Evidencias de Conocimiento: EV1 Contesta preguntas acerca de analítica de datos	IDENTIFICA LAS CARACTERÍSTICAS DE BIG DATA PARA SU CONTEXTUALIZACIÓN DE ACUERDO AL ENTORNO.	IEV1. Cuestionario
Evidencias de Desempeño. EV2 Desarrolla el Caso de estudio de analítica de datos.	SELECCIONA LA HERRAMIENTA PARA LA LIMPIEZA DE DATOS DE ACUERDO A LA MUESTRA Y REQUERIMIENTO DEL USUARIO.	IEV2. Lista de verificación
Evidencias de Producto: EV3 Construye y expone un plan de creación de la hipótesis, determinar fuentes de información, herramientas a utilizar, métodos de recolección, reducción de la dimensional para construir la muestra de un caso propuesto.	CLASIFICA Y ORDENA LAS VARIABLES SEGÚN LOS REQUERIMIENTOS.	IEV3. Lista de verificación



5. GLOSARIO DE TÉRMINOS

- **Población:** Grupo completo de individuos u objetos que constituyen la base de interés para un estudio estadístico. Es el conjunto de todos los elementos que cumplen una determinada característica que deseamos medir y estudiar.
- **Muestra:** parte representativa de una población. Es todo subconjunto de una población sobre la que se va a realizar el estudio.
- **El número** de elementos de la muestra se denomina tamaño de la muestra. Individuo En estadística se considera individuo (objeto) a cada uno de los elementos de la población.
- **Carácter:** Cada uno de los aspectos o propiedades que se pueden estudiar en los individuos de una población recibe el nombre de carácter o estadístico. Esto permite clasificar a los individuos. El carácter puede ser cuantitativo si se puede medir o bien cualitativo si no se puede medir, pero se puede comparar.
- **Dato:** valor o forma que asume una variable para un individuo determinado.
- **Espacio Muestral:** Conjunto de todos los posibles resultados de un experimento estadístico.
- **Variable** El conjunto de valores que puede tomar un carácter estadístico se llama variable estadística. Son atributos que poseen o se le pueden asignar a los individuos de una población y que difieren de uno a otro.
- **Cualitativas:** Las que definen cualidades de los individuos; usualmente pueden subdividirse en categorías. Ejemplo: Variable: Sexo. Categorías: M. F.
- **Cualitativa nominal:** Aquellas variables que no siguen ningún orden en específico. Por ejemplo: Colores (Negro, Naranja, Amarillo).
- **Cualitativa ordinal:** Aquellas que siguen un orden o jerarquía. Por ejemplo: Nivel socioeconómico (Alto, medio, bajo).
- **Cualitativa binaria:** En este caso, las variables son solamente dos. Por ejemplo: Si o No, Hombre o Mujer.
- **Indicadoras:** Valores numéricos que se le asignan a las categorías de una Variable Cualitativa. CC – TI
- **Cuantitativas:** cuando los atributos que las definen son cuantificables o medibles numéricamente. Las Variables Cuantitativas pueden ser Discretas o Continuas.
- **Discretas:** Cuando las variables sólo pueden tomar determinados valores, (asumen valores de uno en uno); es decir pueden tomar un número finito o bien infinito numerable de valores
- **Continuas:** Cuando pueden asumir cualquier valor entre dos enteros consecutivos, es decir pueden tomar todos los valores de un intervalo y tan próximos como se quiera.
- **Algoritmo:** Método que describe cómo se resuelve un problema en término de las acciones que se ejecutan y especifica el orden en que se ejecutan estas acciones. Los algoritmos ayudan al programador a planificar un programa antes de su escritura en un lenguaje de programación.
- **Lenguaje de programación:** Es un lenguaje formal creado para expresar procesos que son llevados a cabo, por ejemplo, los computadores; se usan para crear programas de control físico e interno de un computador, quién expresa algoritmos con precisión.
- **Seudocódigo:** Es un lenguaje de descripción de algoritmos que utiliza palabras y



pide indentación. Representación del código que no sigue reglas como un lenguaje de programación. En un pseudocódigo no hay errores de coordinación

- **Compilador:** Es un programa de informática, que traduce un programa escrito de un lenguaje de programación a otro formando un programa igual a la máquina de interpretación. Traduce el código de fuente de un programa en lenguaje de alto nivel a otro de bajo nivel.
- **Intérprete:** Es un productor de resultados en un archivo de fuente en sistemas diferentes. Los programas que se interpretan son más lentos que los compilados por la necesidad de traducir cada programa mientras es ejecutado. Para mejorar el desempeño algunos lenguajes de programación interpretan el código de fuente original en una forma más compacta.
- **Diagrama de flujo:** Es la representación esquemática del algoritmo o proceso, es utilizado en la programación, la economía, los procesos industriales, Cada paso del proceso es representado por un emblema distinto que contiene una breve descripción de la etapa del proceso.
- **Análisis del problema:** El problema tiene que estar definido y comprendido claramente para desarrollar el algoritmo, y para desarrollar el problema necesita recopilar el algoritmo en un lenguaje de programación. Se debe pasar el algoritmo a programa para ver si el programa solucionará el problema.
- **Codificación:** Es el proceso de transformación de un sistema de datos de origen a otro de destino, de esto se segrega como resultado de la información contenida en los datos resultantes que debe ser equivalente a la primera información.
- **Compilación y ejecución:** Es el uso de cualquier editor de escritos corriente que incluye alguna herramienta de provecho para el programador. La compilación consiste en traducir un programa escrito en un lenguaje de programación a otro, la ejecución es iniciar la carga de un programa o archivo ejecutable.
- **Variable.** Se trata de una estructura matemática que puede almacenar cualquier tipo de información, ya sea numérica, alfanumérica, etc... Para entendernos, una variable podría ser como una caja, en la que puedes introducir cualquier cosa (información).

6. REFERENTES BIBLIOGRÁFICOS

- INTRODUCCIÓN A LA ESTADÍSTICA <https://www.iacs.es/>
- Ejercicios interactivos de variables estadísticas <https://www.superprof.es/>
- GLOSARIO ESTADÍSTICO Fuente: Murra y R. Spiegel, Estadística, McGraw-Hill





7. CONTROL DEL DOCUMENTO

	Nombre	Cargo	Dependencia	Fecha
Autor (es)	JOSE FERNANDO GALINDO SUAREZ	INSTRUCTOR	CGMLTI	20/01/2023

8. CONTROL DE CAMBIOS (diligenciar únicamente si realiza ajustes a la guía)

	Nombre	Cargo	Dependencia	Fecha	Razón Cambio del
Autor (es)					

