



FUNCIÓN PÚBLICA



Analítica institucional: el poder de los datos en el sector público

VERSIÓN 1

DIRECCIÓN DE GESTIÓN
DEL CONOCIMIENTO
SEPTIEMBRE DE 2022

**Departamento Administrativo
de la Función Pública**

César Augusto Manrique Soacha

Director

Jesús Hernando Amado Abril

Subdirector

José Fernando Ceballos Arroyave

Secretario General (e)

Jesús Hernando Amado Abril

Director de Gestión del Conocimiento (e)

Armando López Cortés

Director de Participación, Transparencia
y Servicio al Ciudadano (e)

Hugo Armando Pérez Ballesteros

Director de Gestión y Desempeño
Institucional (e)

Hugo Armando Pérez Ballesteros

Director de Desarrollo Organizacional

Francisco Camargo Salas

Director de Empleo Público

Armando López Cortés

Director de Jurídica

Luz Stella Patiño Jurado

Jefe de Oficina de Control Interno

Daniel Canal Franco

Jefe Oficina Asesora de Comunicaciones

Lina Marcela González González

Jefe Oficina Asesora de Planeación (e)

Édgar Alexander Prieto Muñoz

Jefe Oficina de Tecnología de la
Información y las Comunicaciones

Elaborado por:

Carlos Alberto Martínez

MSc, PhD Doctor en Ingeniería

Dirección de Gestión del Conocimiento

Revisión de forma y corrección de estilo

Carolina Mogollón Delgado

Daniela Del Vecchio Rodríguez

Diagramación y diseño

Susana Bonilla Guzmán

Oficina Asesora de Comunicaciones

**Departamento Administrativo
de la Función Pública**

Carrera 6 n.º 12-62, Bogotá, D.C.,
Colombia

Conmutador: 601 739 5656 / 86 -

Fax: 601 739 5657

Web: www.funcionpublica.gov.co

eva@funcionpublica.gov.co

Bogotá, D.C., Colombia.

Tabla de contenido

INTRODUCCIÓN	7
1. CONCEPTOS BÁSICOS DE LA ANALÍTICA INSTITUCIONAL	9
1.1. Tipos de datos	9
1.1.1. Tipos de datos desde la privacidad.....	10
1.1.2. Tipos de datos desde el almacenamiento	11
1.1.3. Tipos de datos desde el tipo de variable	12
1.2. Dimensiones de la calidad de los datos.....	13
1.3. Principales medidas de relación entre los datos	15
1.4. ¿Qué es análisis de datos?	17
1.5. ¿Qué es analítica institucional?.....	18
2. ANALÍTICA INSTITUCIONAL EN LAS ENTIDADES PÚBLICAS	21
2.1. ¿Cuál es la importancia de la toma de decisiones basada en datos para las entidades públicas?.....	21
2.2. ¿Cuáles son los pilares de la analítica institucional?	23
2.3. Tipos de analítica institucional.....	24
2.3.1. Ejemplo de analítica descriptiva y diagnóstica	26
2.3.2. Ejemplo de analítica predictiva	27
2.3.3. Ejemplo de analítica prescriptiva.....	28
2.4. Principales productos de la analítica institucional.....	29

3. CICLO PARA IMPLEMENTAR LA ANALÍTICA INSTITUCIONAL	33
3.1. Pasos para implementar la analítica institucional	33
3.1.1. Identificar la necesidad a partir de las preguntas o proceso a impactar	33
3.1.2. Identificar las fuentes de datos, calidad y transformaciones necesarias: analítica descriptiva y diagnóstica.....	34
3.1.3. Identificar cómo se relacionan los conceptos de la entidad con los datos e indicadores existentes	36
3.1.4. Generar y validar los productos de analítica	36
3.1.5. Validar y evaluar los resultados obtenidos	48
CONCLUSIONES	49
ANEXOS	51
Anexo A. Ejemplo de arquitectura del flujo de datos para la analítica institucional	51
Anexo B. Minería de datos y big data (datos masivos)	64
Anexo C. Tipos de visualizaciones de datos.....	70

Tabla de figuras

Figura 1. Correlación entre la tasa de divorcios en Maine y el consumo per cápita de margarina	1 6
Figura 2. Pilares de analítica institucional.....	2 2
Figura 3. Tipos de analítica institucional	2 4
Figura 4. Ilustración del proceso de analítica descriptiva y diagnóstica	2 6
Figura 5. Ilustración del proceso de analítica predictiva	2 7
Figura 6. Ilustración del proceso de analítica prescriptiva	2 8
Figura 7. Análisis de embarazos adolescentes (panorama general) para el municipio de Medellín	4 4



Introducción

El Departamento Administrativo de la Función Pública presenta el documento **Analítica institucional: el poder de los datos en el sector público** en el marco del modelo integrado de planeación y gestión (MIPG), cuya sexta dimensión es “Gestión del conocimiento y la innovación”, que se desarrolla a partir de 4 ejes: generación y producción; herramientas para uso y apropiación; analítica institucional; y cultura de compartir y difundir. En este documento nos enfocaremos en el eje de analítica institucional, cuyo objetivo es que las entidades adquieran, procesen y analicen sus datos para fortalecer su gestión y desempeño a través de la toma de decisiones basada en evidencia.

Para poder adoptar la analítica institucional es necesario contar con mecanismos que permitan traducir los datos en información de valor a través de herramientas, metodologías y procedimientos replicables que identifican patrones y comportamientos en los datos. Para abordar estos temas, estructuramos este documento tres capítulos, así: el primer capítulo introduce los conceptos básicos de la analítica institucional; en el segundo capítulo se trata la analítica institucional en las entidades públicas, que incluye los pilares y los tipos y principales productos de la analítica institucional; en el tercer capítulo se muestra el ciclo para implementar la analítica institucional a partir de cada uno de los pasos sugeridos; y, finalmente, se incluye una sección de conclusiones generales.

Adicionalmente, encontrarán tres anexos que complementan el ciclo de implementación propuesto, que son: Anexo A. *Ejemplo de arquitectura del flujo de datos para la analítica institucional*, Anexo B. *Minería de datos y big data* (datos masivos) y Anexo C. *Tipos de visualizaciones de datos*.



1. Conceptos básicos de la analítica institucional

Este apartado introduce los conceptos básicos de la analítica institucional y la importancia de esta práctica en la toma de decisiones. Para ello, describiremos la diferencia entre dato e información y expondremos los distintos tipos de datos existentes, las dimensiones de calidad de los datos y las principales medidas de relación entre estos.

1.1. Tipos de datos

Según Davenport & Prusak (1998), un dato es un conjunto discreto de factores objetivos sobre un hecho real, es decir, un valor sobre algún hecho, por ejemplo:

- 54
- Masculino
- 2022
- 2021-01-05
- MIPG

Estos datos pueden estar almacenados en las celdas de una hoja de cálculo (Excel), formatos en físico o en una base de datos, entre otros.

Por otra parte, la información corresponde a datos con contexto (Davenport & Prusak, 1998, p. 3), es decir, que tienen significado en un ámbito específico. A partir de los datos anteriores se puede obtener la siguiente información:

- Edad: 54
- Sexo: masculino
- Año actual: 2022
- Fecha de actualización: 2021-01-05
- Fuente de información: MIPG

Para que las entidades tomen decisiones basadas en evidencia, es necesario que recolecten, almacenen y transformen los datos en información de valor. Lo anterior supone costos de infraestructura (almacenamiento, licenciamiento de *software*, etc.) y esfuerzo de transformación y visualización de los datos (personas, tiempo de procesamiento y energía, entre otros).

Los datos pueden ser catalogados desde distintos enfoques, por ejemplo: privacidad, almacenamiento o tipo de variable.

1.1.1. Tipos de datos desde la privacidad

En el artículo 3° de la Ley 1266 de 2008 se define que son:

... **e) Dato personal.** Es cualquier pieza de información vinculada a una o varias personas determinadas o determinables o que puedan asociarse con una persona natural o jurídica. Los datos impersonales no se sujetan al régimen de protección de datos de la presente ley. Cuando en la presente ley se haga referencia a un dato, se presume que se trata de uso personal. Los datos personales pueden ser públicos, semiprivados o privados;

f) Dato público. Es el dato calificado como tal según los mandatos de la ley o de la Constitución Política y todos aquellos que no sean semiprivados o privados, de conformidad con la presente ley. Son públicos, entre otros, los datos contenidos en documentos públicos, sentencias judiciales debidamente ejecutoriadas que no estén sometidos a reserva y los relativos al estado civil de las personas;

g) Dato semiprivado. Es semiprivado el dato que no tiene naturaleza íntima, reservada, ni pública y cuyo conocimiento o divulgación puede interesar no sólo a su titular sino a cierto sector o grupo de personas o a la sociedad en general, como el dato financiero y crediticio de actividad comercial o de servicios a que se refiere el Título IV de la presente ley.

h) Dato privado. Es el dato que por su naturaleza íntima o reservada solo es relevante para el titular...

Además, el artículo 5° de la Ley 1581 de 2012 de protección de datos personales determina como categoría especial de datos:

... **Datos sensibles.** (...) se entiende por datos sensibles aquellos que afectan la intimidad del Titular o cuyo uso indebido puede generar su discriminación, tales como aquellos que revelen el origen racial o étnico, la orientación política, las convicciones religiosas o filosóficas, la pertenencia a sindicatos, organizaciones sociales, de derechos humanos o que promueva intereses de cualquier partido político o que garanticen los derechos y garantías de partidos políticos de oposición, así como los datos relativos a la salud, a la vida sexual y los datos biométricos...

1.1.2. Tipos de datos desde el almacenamiento

Al almacenar los datos, se puede considerar que existen tres tipos, los cuales a partir de Naeem, T. (2020), pueden entenderse como:

- **Estructurados:** son datos que están organizados en una estructura definida, por ejemplo: hojas de cálculo o tablas en bases de datos, entre otros.
- **No estructurados:** son datos libres sin un formato específico, por ejemplo: videos, audios o textos.
- **Semiestructurados:** son una combinación de los dos anteriores, en el cual existe un orden general pero los datos

pueden ser no estructurados, por ejemplo: hojas de Excel con imágenes y texto libre en ciertas columnas, entre otros.

1.1.3. Tipos de datos desde el tipo de variable

De acuerdo con Salazar y del Castillo (2018), los datos se pueden clasificar según el tipo de variable, que es cualitativa o cuantitativa (pp. 15-16).

1. VARIABLES CUALITATIVAS

“Las variables cualitativas se refieren a características o cualidades que no pueden ser medidas con números” (Salazar, C. y Del Castillo, P. 2018 p. 16), estos autores identifican dos tipos de variables:

- **Nominales:** variables no numéricas que no tienen orden, ejemplo: sexo, nacionalidad, país o identidad de género, entre otros.
- **Ordinales:** variables no numéricas que tienen orden, ejemplo: estrato socioeconómico, estado de un proceso o escala de satisfacción, entre otros.

2. VARIABLES CUANTITATIVAS

“Una variable cuantitativa es la que se expresa mediante un número, por tanto, se pueden realizar operaciones aritméticas con ella.” (Salazar, C. y Del Castillo, P. 2018 p. 17), estos autores identifican dos tipos:

- **Discretas:** son aquellas variables que toman valores aislados o enteros sin valores intermedios¹, es decir, que no existe un valor entre uno y otro, por ejemplo: número total de servidores en una entidad (no hay una persona que valga por media).

1 Se entenderá por valores intermedios a los números decimales, es decir, que constan de una parte entera y una decimal, separadas por un punto o por una coma.

- **Continuas:** son aquellas variables que pueden tomar valores intermedios, es decir, que puede existir un valor decimal entre uno y otro, por ejemplo: promedio de la altura de las personas en un municipio o promedio de la edad de los servidores públicos, entre otras.

1.2. Dimensiones de la calidad de los datos

La calidad de los datos es importante para una entidad porque le permitirán entender la realidad del área, proceso, ciudadano/a o territorio, entre otros. Es importante recalcar que, si nuestros datos son de mala calidad, obtendremos resultados sesgados e inadecuados que propiciarán la toma de malas decisiones. Por ejemplo, si una entidad cuenta con pocos datos de sus grupos de valor como la edad, el género, el lugar de origen, etc., la llevaría a analizar a todos como iguales, desconociendo el enfoque diferencial y, posiblemente, definirá acciones generalistas que no aplican a todos los grupos poblacionales, lo que disminuirá la efectividad de sus acciones.

Cabe señalar que la calidad de los datos no es una medida única, esta depende de las características propias del dato. En este documento usaremos las dimensiones para medir la calidad de los datos que sintetiza Ortega (s.f.) con base en la *Data Administration Management Association* (DAMA) (2017), estas son: exactitud, completitud, consistencia, pertinencia temporal, unicidad y validez:

- **Exactitud** (*Accuracy*): Se mide el grado en el que los datos representan correctamente el objeto del mundo real o un evento que se describe. (...)
- **Completitud** (*Completeness*): El grado en el que el dato tiene el valor esperado y cumple con los requerimientos marcados. (...)

- **Consistencia** (*Consistency*): Mide si los datos están libres de contradicción y tienen coherencia lógica, de formato o temporal. (...)
- **Pertinencia temporal** (*Timeliness*): Mide el grado en que los datos están disponibles cuando se requieren. (...)
- **Unicidad** (*Uniqueness*): Cada dato es único. Con esta dimensión se busca corregir la duplicidad inesperada en nuestros dataset. (...)
- **Validez** (*Validity*): Medir si un valor se ajusta a una regla de negocio o a un estándar preestablecido en cuanto a formato, tipo de dato, valores posibles o rangos especificados. (...)

Algunos ejemplos de las dimensiones de calidad de los datos en el contexto de las entidades públicas son: exactitud, que la dirección de envío del predial en la base de datos corresponda con la dirección real; completitud, si consideramos que los de contacto de una persona están completos cuando contamos con: nombre(s), primer apellido, segundo apellido, número de identificación, correo electrónico y dirección; consistencia, no hay consistencia si tenemos registro de pagos del impuesto predial sin que se haya liquidado el valor a pagar; pertinencia temporal, para la asignación de un subsidio la entidad debe contar con el listado de disponibilidad de cupos en ese momento; unicidad, no hay unicidad si la misma persona aparece en dos o más registros distintos en la misma base de datos (Juan F Hoyos C, Juan Felipe Hoyos Cruz, Felipe H, etc.); y validez, no hay validez cuando en una mesa votación aparece registro de voto de una persona fallecida.

De acuerdo con lo anterior, las dimensiones son atributos que le servirán a las entidades como una guía para seleccionar los conjuntos de datos apropiados para el análisis y entender el nivel de confiabilidad que estos tienen al momento de concluir algo a partir de ellos. Cabe resaltar que el nivel de calidad de los datos requerido debe estar alineado a la necesidad de análisis de la

entidad y, aunque lo recomendado es que haya un 100 % de calidad de los datos, esto no significa que no se pueda llevar a cabo con porcentajes inferiores.

1.3. Principales medidas de relación entre los datos

Una vez la entidad considere que los datos que tiene son de calidad, sugerimos que analice las relaciones que existen entre ellos, que pueden corresponder a su comportamiento en el tiempo: tendencia, ciclo y estacionalidad; o a su relación con otras variables: correlación y causalidad.

Frente al comportamiento en el tiempo, una serie de tiempo es "... is a set of observations x_t , each one being recorded at a specific time t ..." (Brockwell & Davis, 2016, p. 1), es decir, a un conjunto de observaciones registradas en un momento específico, que se pueden analizar a partir de sus componentes:

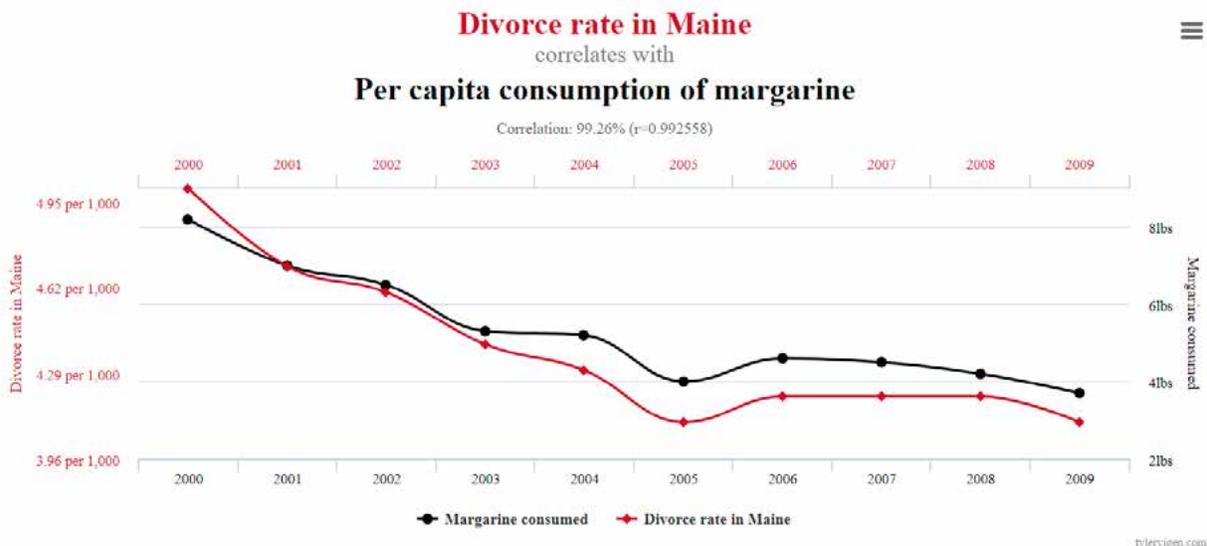
- **Tendencia:** corresponde con el comportamiento o movimiento de una variable (aumenta o disminuye) en un periodo de tiempo definido. Por ejemplo: el promedio de hijos por mujer ha disminuido desde 1985 a 2018 (fuente: Fuente: DANE – Proyecciones de población derivadas del CNPV 2018), es decir, la tendencia es a disminuir.
- **Ciclo:** corresponde con aquellos comportamientos repetitivos en el tiempo con algún tipo de frecuencia mayor a un año, es decir, que se repite cada X tiempo. Por ejemplo: el ciclo anual de lluvias.
- **Estacionalidad:** corresponde con ciertos valores cercanos entre sí que oscilan de una manera constante en un periodo menor a un año. Por ejemplo: consumo energético y de agua en una entidad o precipitación de agua regional.

En cuanto a la relación de una variable con otras variables, la correlación y causalidad pueden entenderse según Tshilidzi (2015) como:

- **Correlación:** medida o grado de relación entre dos variables, es decir, que al aumentar o disminuir una la otra tiene un comportamiento similar. Por ejemplo, cuando llueve aumentan los casos de personas resfriadas y también aumentan los accidentes de tránsito por deslizamiento del vehículo en la vía, es decir, hay una correlación porque, cuando aumenta una variable (número de personas resfriadas), aumenta la otra variable (número de accidentes de tránsito por deslizamiento del vehículo en la vía) en el mismo periodo de tiempo.
- **Causalidad:** describe la relación entre causas y efectos. En nuestro ejemplo anterior, el resfriado no es la causa de los accidentes de tránsito por deslizamiento del vehículo en la vía, sino que existe un factor común que genera la causalidad, en este caso es la lluvia.

Durante el proceso de análisis debemos prestar atención a las relaciones entre las variables para saber si realmente corresponden a un principio de causalidad o simplemente existe correlación entre estas. Para ilustrar esta idea, en el siguiente ejemplo (ver Figura 1) se muestra una alta correlación entre la tasa de divorcios en el Estado de Maine, EE. UU., y el consumo per cápita de margarina, pero, a pesar de su alto grado de correlación, no se puede concluir que haya una relación de causalidad.

Figura 1. Correlación entre la tasa de divorcios en Maine y el consumo per cápita de margarina



Nota. Imagen tomada de tylervigen.com

1.4. ¿Qué es análisis de datos?

El análisis de datos puede definirse de distintas maneras; sin embargo, todas ellas coinciden en que es una capacidad organizacional que permite procesar nuestros datos para transformarlos en información de valor para la toma de decisiones. Los datos son la materia prima para responder a las preguntas de la entidad; en otras palabras, podemos decir que los datos son la harina para producir galletas, pero la harina también puede ser empleada para hacer pasteles, tortas y otras preparaciones, de modo que es el propósito el que determina su uso.

Entonces, debemos preguntarnos: ¿cómo utilizamos los datos para dar respuesta a las preguntas relevantes del día a día de nuestra entidad? y ¿cómo estas preguntas son resueltas por medio del análisis de datos? Para hacer análisis de datos debemos definir qué vamos a analizar (plan, programa, proceso, necesidad, etc.), con esto, recolectamos los datos disponibles y los cruzamos a partir de

distintas fuentes (archivos, hojas de cálculo, bases de datos) para agruparlos en conceptos comunes, unificarlos y definir indicadores que permitan entender la relación entre los datos (qué pasó y por qué pasó).

1.5. ¿Qué es analítica institucional?

La analítica institucional es un eje de la dimensión “Gestión del conocimiento y la innovación” del modelo integrado de planeación y gestión (MIPG) que busca “Profundizar en los análisis de la información y los datos que genera la entidad para fortalecer el conocimiento de su desempeño y de su propósito fundamental.” (Función Pública, 2021, p. 102) a través de cuatro criterios:

- **Diagnóstico de datos:** el MIPG lo define como la “Identificación, organización y actualización de los datos producidos por la entidad” (Función Pública, 2021, p.102), es decir, entender qué datos tenemos al interior de nuestras entidades. Por ejemplo, hacer un diagnóstico de los datos, de sus procesos, los grupos de valor, las acciones o políticas que ejecuta en territorio, entre otros. Los datos pueden estar almacenados en físico (planillas, documentos impresos, carpetas y libros) o en algún medio digital (hojas de cálculo, documentos digitales, bases de datos, aplicaciones internas o de terceros). El propósito de reconocer los datos producidos por la entidad es definir la situación actual y organizarlos para que sean útiles y estén actualizados.
- **Análisis de datos:** el MIPG lo define como “Profundizar en los análisis de datos internos y externos, haciendo cruces que permitan extraer nuevas perspectivas de los datos y de la información” (Función Pública, 2021, p.102). Para ello es necesario que conozcamos y entendamos los datos; por eso es importante tener en cuenta que una sola fuente de datos (una hoja de cálculo, un documento o una tabla en una base

de datos) suele ser insuficiente, por lo que debemos cruzarla con otras que nos permitan obtener una visión más amplia de lo que está sucediendo. Por ejemplo, si tengo el listado de todos los niños con malnutrición identificados en los últimos seis meses, podría profundizar el conocimiento indagando cuáles de ellos aún no han sido cubiertos por programas institucionales para realizar un seguimiento e incorporarlos.

- **Visualizaciones:** según el MIPG las visualizaciones se usan para "... contar historias a través de los datos". Visualizar datos no solo implica usar tablas y hojas de cálculo, también incluye imágenes o gráficos, entre otros, que permitan entender el comportamiento de los datos y concluir sobre ellos en poco tiempo y de forma fácil.
- **Tableros de indicadores:** según el MIPG (2021) los tableros de indicadores son una forma de visualizar datos que se usan para monitorear la gestión de la entidad, esto es clave porque lo que no se mide es invisible para la gestión institucional. Para ello, es necesario definir indicadores con valores de referencia que nos permitan conocer si estamos haciendo bien nuestra gestión (monitoreo periódico) y, en caso contrario, asociar acciones de mejoramiento (gestión diaria). Por ejemplo, podríamos definir que nuestra meta es: cero niños menores de cinco años con malnutrición en nuestro municipio. Para cumplir la meta definiremos un plan de atención a un plazo de seis meses para atender a 200 niños mes a mes. Nuestro indicador será: número de niños atendidos mes a mes con un valor de referencia (meta) de 200. Para el monitoreo de la gestión nuestro indicador será visualizado por los líderes del programa al interior de la entidad a través de un tablero de indicadores que permita ver la evolución histórica, los valores actuales y las alertas de incumplimiento, si las hay.



2. Analítica institucional en las entidades públicas

2.1. ¿Cuál es la importancia de la toma de decisiones basada en datos para las entidades públicas?

La analítica de datos es importante porque permite:

- **Tomar decisiones con evidencia y no por suposición:** tomar decisiones con base en experiencia particular, información parcial o sin ser corroborada con datos propicia la toma de decisiones erróneas. Por ejemplo: no viajar en avión debido a que son inseguros porque ocurrió un accidente esta semana, desconociendo que, según la Flight Safety Foundation, el avión es considerado el medio de transporte más seguro por número de accidentes anuales.
- **Ser más efectivos, precisos y eficientes:** si conocemos el resultado de las acciones que tomamos (medición y monitoreo) podemos seleccionar aquellas que fueron más efectivas y eficaces para implementarlas en nuestra entidad, mientras que podemos limitar aquellas que no han dado los resultados esperados.
- **Disminuir los costos:** realizar acciones que no son efectivas una y otra vez conlleva costos innecesarios, el no conocer el público de interés para un programa o una política podría generar que se implemente de forma inadecuada o

no alineada a sus expectativas. Por lo que medir, analizar y conocer qué sucede al interior de entidad y los distintos públicos que la conforma permite disminuir los costos.

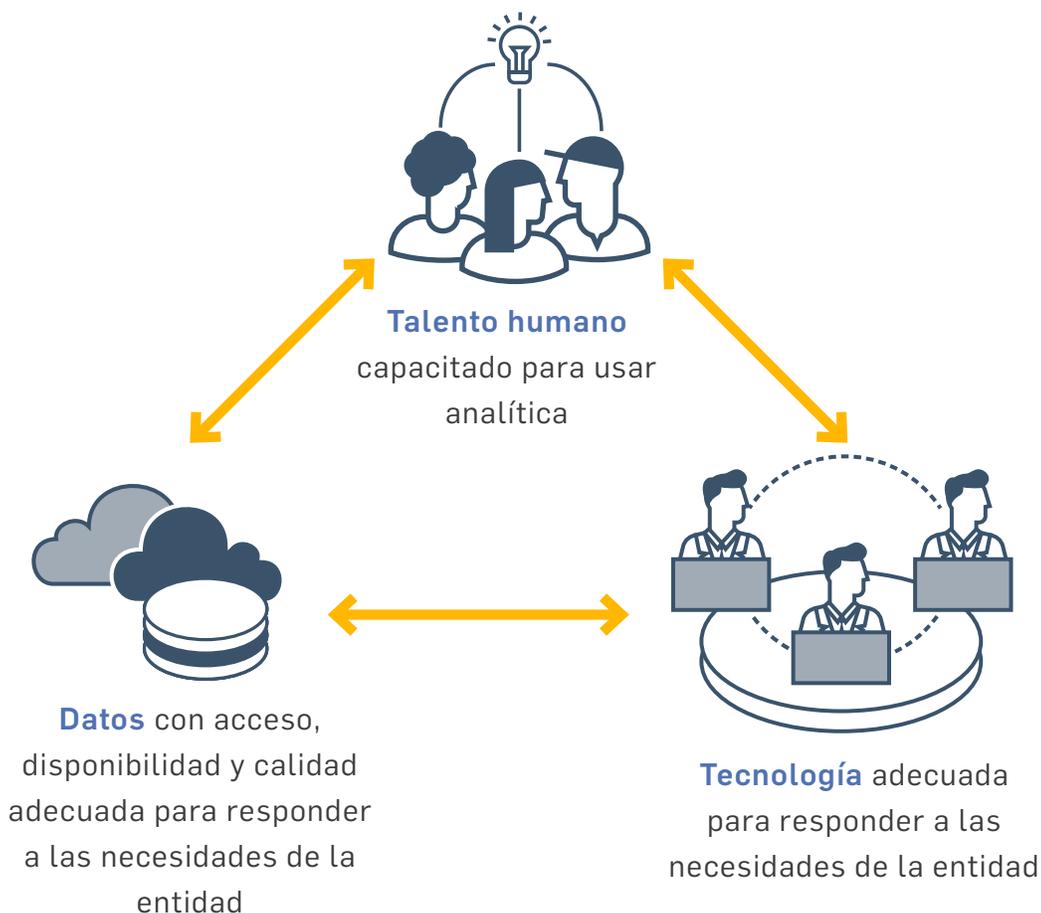
- **Disminución los riesgos:** para disminuir los riesgos asociados a una práctica inadecuada o la omisión y falta de control de alguno de los procesos, es importante que la entidad conozca, articule y analice sus datos, por ejemplo, sobre el nivel de endeudamiento, la ejecución de proyectos y el impacto de estos, entre otros. Lo anterior se podría entender como un adecuado proceso de análisis de datos.
- **Disminuir los errores:** si como entidad puedo conocer los cuellos de botella y cuáles son los errores que estoy cometiendo (por acción u omisión) puedo realizar un proceso de identificación de causales y trabajarlas de forma adecuada, por lo que la medición y el análisis es crucial para disminuir los errores y las brechas de conocimiento.
- **Identificar oportunidades no evidentes:** muchas veces podemos tener oportunidades de mejora en nuestra entidad y las desaprovechamos porque no nos damos cuenta de que existen. Por ejemplo, una entidad automatiza el apagado de luces en las oficinas para después de las 6:00 p.m., pero identifica que está desaprovechando ahorros en energía en ciertos periodos en los que no hay personas laborando en sus puestos de trabajo, por ejemplo, entre las 12:30 p. m. y la 1:30 p.m. (hora de almuerzo), de modo que incluyó esa franja horaria en el sistema de apagado y encendido automático de luces de la entidad.

El análisis de datos debe traducirse en mejores resultados que puedan ser validados en el día a día de nuestra entidad, el territorio y por los ciudadanos.

2.2. ¿Cuáles son los pilares de la analítica institucional?

Para que una entidad pueda llevar a cabo la analítica institucional debe contar no solo con datos, sino también con talento humano capacitado y tecnología alineada a las necesidades de la entidad. Es así como se propone abordar el desarrollo de la analítica al interior de nuestras entidades a través de tres pilares: personas, datos y tecnología, (ver Figura 2).

Figura 2. Pilares de analítica institucional



Nota. Diseñado por el Departamento Administrativo de la Función Pública (2022).

Como se muestra en la Figura 2, los 3 pilares son necesarios porque si faltara el talento humano, sería como tener una nave espacial sin

nadie que la pilotee (dado que necesitamos a alguien que conozca su funcionamiento y dé dirección al timón); si faltaran los datos, nos faltarían los insumos para los análisis; si faltara la tecnología, no podríamos responder en un tiempo adecuado.

2.3. Tipos de analítica institucional

La analítica comprende distintos tipos de análisis de datos. El documento Aprovechamiento de datos para la toma de decisiones en el sector público (2021) del Departamento Nacional de Planeación (DNP) plantea que los análisis de datos se pueden agrupar en análisis descriptivo, de causalidad, predictivo y prescriptivo (p. 27). Para este documento adaptaremos los tipos de análisis definidos por el DNP así:

- **Analítica descriptiva (análisis descriptivo):** da respuesta a qué pasó o está pasando.
- **Analítica diagnóstica (análisis de causalidad):** da respuesta a por qué pasó o está pasando.
- **Analítica predictiva (análisis predictivo):** da respuesta a qué puede pasar en el futuro.
- **Analítica prescriptiva (análisis prescriptivo):** da respuesta a cuál es la mejor opción.

Si bien el ideal es llevar a cabo los cuatro tipos de analítica en la entidad, primero es necesario que sepa qué está sucediendo con respecto a un problema, necesidad, proceso, plan, etc., (analítica descriptiva) para entender por qué sucede (analítica diagnóstica). A partir de lo anterior, la entidad podrá predecir el comportamiento futuro del objeto de análisis (analítica predictiva) y así identificar cuál será la decisión que mejor se ajuste a los criterios definidos por la entidad (analítica prescriptiva).

En la Figura 3 se muestran los cuatro tipos de analítica por nivel de complejidad, las preguntas a las que responde cada una y un ejemplo:

Figura 3. Tipos de analítica institucional

DESCRIPTIVA	DIAGNÓSTICA	PREDICTIVA	PRESCRIPTIVA
¿Qué pasó o está pasando?	¿Por qué pasó o está pasando?	¿Qué pasará?	¿Cuál sería mi mejor decisión?
Ayer y hoy ha llovido.	Gran evaporación de fuentes hídricas aledañas e inicio del fenómeno de La Niña.	Mañana muy posiblemente llueva.	Para ir al centro del municipio sin mojarme al menor costo posible debería salir 30 minutos antes e ir en bicicleta.

Nota. Adaptado por el Departamento Administrativo de la Función Pública (2022) con base en el documento Aprovechamiento de datos para la toma de decisiones en el sector público del Departamento Nacional de Planeación, 2021, p. 28.

Como se ve en la Figura 3, cada tipo de analítica responde a una necesidad de datos, por lo que hacer un solo tipo de analítica limita el proceso de toma de decisiones. Por ello, para implementar cualquier tipo de analítica es necesario:

- **Definir la pregunta para resolver:** esta es la base de cualquier tipo analítica de datos, para ello es necesario identificar el propósito de la pregunta, conocer qué está sucediendo y cuál es el proceso por impactar.
- **Retroalimentar los procesos:** los resultados de la analítica de datos no dan soluciones por sí solos, de modo que es necesario que estos repercutan en la realidad de la entidad (planes, programas, proyectos y necesidades, entre otros) y no queden en un ejercicio meramente académico. Por ejemplo, el resultado del análisis de datos arrojó que la

congestión vehicular se debía a la configuración actual de la red semafórica, se encontró una configuración óptima que puede disminuir el tiempo de desplazamiento en promedio en 30 minutos, de modo que la entidad debe decidir si implementará o no los cambios propuestos en la red semafórica.

A continuación, abordaremos cada tipo de analítica a través de un ejemplo.

2.3.1. Ejemplo de analítica descriptiva y diagnóstica

Un ejemplo de proceso de analítica descriptiva y diagnóstica puede ser la toma del registro de lluvias por día de los últimos 30 días para identificar los días en los que ha llovido y en los que no (análisis histórico), luego, se propone la elaboración de un cuadro comparativo que evidencie si las lluvias están aumentando o disminuyendo y así establecer una tendencia, posteriormente, esta información se cruza o complementa con el registro histórico de temporadas de lluvia y de fenómenos de El Niño o La Niña en el mismo periodo de tiempo para identificar posibles correlaciones y, de esa manera, plantear una conclusión. Para este caso podríamos decir, por ejemplo, que en los últimos siete días ha llovido (analítica descriptiva) debido al inicio del fenómeno de La Niña (analítica diagnóstica).

Con base en el ejemplo anterior, en la Figura 4 encontrará una ilustración del proceso de analítica descriptiva y diagnóstica:

Figura 4. Ilustración del proceso de analítica descriptiva y diagnóstica



Nota. Adaptado por el Departamento Administrativo de la Función Pública (2022) con base en el documento Aprovechamiento de datos para la toma de decisiones en el sector público del Departamento Nacional de Planeación, 2021, p. 28.

2.3.2. Ejemplo de analítica predictiva

El análisis de los datos históricos nos permite identificar si estos pertenecen a un grupo, clase o segmento, o predecir, con cierto grado de certeza, si ocurrirá o no evento. Dando continuidad al ejemplo anterior sobre el registro de lluvias, en el que se identificó que en los últimos siete días llovió debido al inicio del fenómeno de La Niña. Con esa información se podría construir un modelo matemático (ver sección 2.5. Principales productos de la analítica institucional) que estime la probabilidad de lluvia del siguiente día, para nuestro ejemplo podemos suponer que arrojó una estimación del 98%, esto significa que hay una alta probabilidad de lluvia para el día de mañana.

Figura 5. Ilustración del proceso de analítica predictiva



Registro de lluvias por día de los últimos 30 días.

Registro histórico de las temporadas de lluvia y de los fenómenos de El Niño o La Niña.

Probabilidad de que llueva mañana dado que ha llovido en los últimos 7 días.

MUY POSIBLEMENTE LLUEVA
(PROBABILIDAD DEL 98%).

Nota. Adaptado por el Departamento Administrativo de la Función Pública (2022) con base en el documento Aprovechamiento de datos para la toma de decisiones en el sector público del Departamento Nacional de Planeación, 2021, p. 28.

2.3.3. Ejemplo de analítica prescriptiva

De acuerdo con lo anterior, si sabemos que la probabilidad de lluvia es muy alta, debemos encontrar la mejor opción² para el día de mañana frente al siguiente escenario: “Debo ir a la alcaldía de mi municipio en el menor tiempo posible y sin mojarme, teniendo en cuenta que vivo a 15 km”. Para dar una respuesta podríamos aplicar un modelo que seleccione el mejor medio de desplazamiento al municipio que minimice el costo y evite que me moje. La mejor opción encontrada y que responde a todos los criterios mencionados es: caminar al punto de parada del bus con sombrilla y tomar la ruta X en dirección a la alcaldía.

2 Cuando hablamos de “la mejor opción” depende de cada situación y de la pregunta a resolver, algunas veces no es solo un criterio como el costo, si no varios como el tiempo, esfuerzo y disponibilidad, entre otros.

Figura 6. Ilustración del proceso de analítica prescriptiva



Nota. Adaptado por el Departamento Administrativo de la Función Pública (2022) con base en el documento Aprovechamiento de datos para la toma de decisiones en el sector público del Departamento Nacional de Planeación, 2021, p. 28.

2.4. Principales productos de la analítica institucional

Los principales productos de analítica institucional en nuestras entidades son:

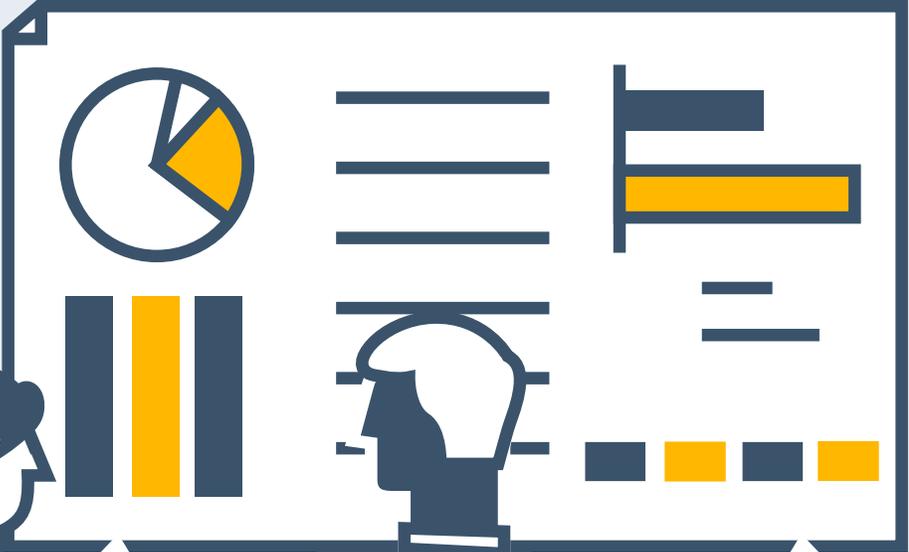
- Modelos de datos: pueden entenderse como una manera de organizar los datos de forma estructurada que permita hacer consultas y análisis sobre los datos de la entidad y responda a las necesidades de información de la entidad. Por ejemplo, bases de datos con los datos de radicados que le permitan a un servidor(a) consultarla y saber, por ejemplo, cuáles

radicados están en proceso, cuáles vencidos, a qué área está asignado o revisar el contenido de un radicado específico, entre otros.

- **Indicadores clave de desempeño (KPI):** Kerzner, H. (2017) define KPI (key performance indicator) como aquel indicador clave de desempeño o rendimiento que debe estar alineado con los objetivos del proceso, área o entidad, entre otros, y permite identificar cuál su estado de desempeño o rendimiento. Los KPI se suelen definir porcentualmente para compararlos entre distintos procesos, áreas, zonas geográficas, etc., asociándoles un nivel o valor esperado (meta) y algún mecanismo de validación. Por ejemplo, el KPI es: porcentaje de cumplimiento diario en la atención en ventanilla. La meta para este KPI puede ser que el porcentaje diario de atención en ventanilla esté por encima del 80 %. El porcentaje de cumplimiento de la meta se puede validar dividiendo el total de personas atendidas en el día sobre el total de personas que solicitaron turno para esa ventanilla. Esto le permite comparar el cumplimiento de la meta durante un periodo de tiempo para determinar si es necesario aumentar o disminuir la capacidad.
- **Procesos automáticos de cálculo de indicadores:** se refiere a un proceso ejecutado a través de un programa de software que calcula indicadores de forma automática, actualizada y confiable en un periodo de tiempo definido (diario, semanal, mensual, etc.) sin la necesidad de intervención humana. Por ejemplo, la entidad se propone atender a 200 personas en una semana, para ello puede desarrollar un procedimiento o programa de software que consolide de forma automática el número de personas atendidas en un rango de tiempo y con esa información calcular el porcentaje de cumplimiento del indicador asociado a la meta esperada, este puede ser porcentaje de personas atendidas en X tiempo.
- **Reportes y tableros de control:** son informes o visualizaciones que permiten entender los datos de una

manera fácil para el análisis y toma de decisiones en el día a día. Recordemos que el valor de un reporte o de un tablero de control radica en los análisis y decisiones que se generan a partir de ellos. Por ejemplo, tableros de control que nos permitan identificar la evolución de la mal nutrición en los niños menores de 5 años en mi territorio para definir la inversión y los planes, programas o proyectos institucionales por implementar.

- Servicio de consulta de datos específicos: son aquellos sistemas o programas de software que le permiten a cualquier servidor(a) de la entidad consultar, seleccionar y filtrar datos específicos sin necesidad de requerimientos a otras áreas. Por ejemplo, que una alcaldía desarrolle un servicio de consulta de datos en su página web que les permita a sus servidores(as) obtener el listado de vehículos matriculados en su municipio en un periodo determinado.
- Modelos matemáticos: se pueden entender como la combinación de operaciones matemáticas que me permiten representar la relación entre dos o más variables. Por ejemplo, si una entidad quisiera saber el consumo de gasolina de sus vehículos en el próximo mes podría hacerlo a través de un modelo matemático en el que se calcule el promedio de consumo del mes anterior.



3. Ciclo para implementar la analítica institucional

3.1. Pasos para implementar la analítica institucional

Para implementar la analítica institucional en el quehacer diario de las entidades del Estado recomendamos seguir los siguientes pasos:

1. Identificar la necesidad a partir de las preguntas o proceso a impactar.
2. Identificar las fuentes de datos, calidad y transformaciones necesarias.
3. Identificar cómo se relacionan los conceptos de la entidad con los datos e indicadores existentes.
4. Generar y validar los productos de analítica.
5. Validar y evaluar los resultados obtenidos.

3.1.1. Identificar la necesidad a partir de las preguntas o proceso a impactar

El primer paso para implementar la analítica institucional es saber cuál es la necesidad que se pretende resolver (procesos, planes, programas, proyectos, etc.), por tanto, las preguntas analíticas son el punto de partida para identificar las necesidades y problemáticas

de la entidad (Departamento Nacional de Planeación, 2020 p. 87). Para plantear estas preguntas recomendamos:

1. Entender el propósito de la pregunta: para ello podemos plantearnos varias veces consecutivas el “para qué”.
2. Identificar si corresponde a una pregunta sobre:
 - **Análisis de datos:** necesito una base de datos que..., un reporte que..., un tablero que...
 - **Futuro:** ¿Cuál es la posibilidad/probabilidad de...?, ¿Quiénes serán los ciudadanos que...? ¿Cuál es el comportamiento esperado de...?
 - **Simulación:** ¿Qué pasaría si...? ¿Qué lograríamos al cambiar...?
 - **Optimización:** ¿Cuál es la solución a menor costo? ¿Cuál debería ser la mejor secuencia de atención de necesidades de mi ciudadano? etc.
3. Verificar que la pregunta esté alineada con el propósito de la entidad.
4. Unificar preguntas similares entre áreas de la entidad.

3.1.2. Identificar las fuentes de datos, calidad y transformaciones necesarias: analítica descriptiva y diagnóstica

Para llevar a cabo el ejercicio de analítica institucional debemos identificar las fuentes de datos, calidad y transformaciones necesarias, para ello sugerimos a las entidades tener en cuenta:

- **Las fuentes y tipos de datos:** tener claro dónde están los datos y cuáles son sus tipos (ver apartado 1.1. Tipos de datos).

- **Si los datos corresponden a una serie de tiempo:** identificar si los datos son tomados en periodos regulares de tiempo y si en estos existe tendencia, ciclo o estacionalidad e incluir esta información al momento del análisis (ver apartado 1.3. Principales medidas de relación entre los datos).
- **La causalidad y correlación:** verificar si en los datos seleccionados existe causalidad o correlación e incluir esta información al momento del análisis (ver apartado 1.3. Principales medidas de relación entre los datos).
- **La gestión y el gobierno de los datos:** cumplir los lineamientos definidos por la entidad para la gestión y el uso de los datos. Por ejemplo, identificar cuáles son los roles y responsabilidades de cada uno de los actores involucrados, cuál es el proceso para solicitar información al interior de la entidad, cuáles datos pueden ser de uso público o cuáles tienen restricciones de acceso de acuerdo con el tipo y uso especificados en la Ley de Protección de Datos Personales o Ley 1581 de 2012, entre otros.

Una vez realizado lo anterior, se debe documentar la información generada por medio de un diccionario de datos o similar para su uso por la entidad.

Los datos son solo eso hasta que les damos contexto y los transformamos en información; no debemos solo identificar un dato, debemos hacernos preguntas sobre “el por qué” de ese dato: ¿Qué significa? ¿Por qué ese valor? ¿Es adecuado o no? ¿Cómo llegamos a ese valor? ¿Con cuáles datos se relaciona?, etc.



En caso de que los datos no existan o nuestra entidad no pueda acceder a ellos, se sugiere:

- Proponer un proyecto de adquisición y organización de los datos necesarios.
- Cambiar las preguntas analíticas y comenzar el ciclo de nuevo

3.1.3. Identificar cómo se relacionan los conceptos de la entidad con los datos e indicadores existentes

Los datos e indicadores corresponden a mediciones que se realizan sobre un proceso, grupo de valor, etc., asociados a conceptos propios de la entidad, es decir, con significado para la entidad, por ejemplo: presupuesto de gasto, ciudadano atendido en taquilla, impuesto predial, entre otros. Es así como es necesario entender a qué hacen referencia los datos e indicadores, y si estos últimos corresponden a un indicador clave de desempeño (KPI, por la sigla en inglés de *Key Performance Indicator*), con el fin de incluir este conocimiento al momento del análisis.

3.1.4. Generar y validar los productos de analítica

Una vez definida la necesidad que se pretende resolver y cuáles son los datos e indicadores con los que contamos para ello, procedemos a definir el producto de analítica que puede dar respuesta a dicha necesidad (ver sección 2.5. Principales productos de la analítica institucional).

El desarrollo de productos de analítica puede implicar retos alrededor de: la metodología a emplear, la organización, el almacenamiento de los datos o la adecuada presentación de los resultados, entre otros. A continuación se darán algunas recomendaciones para la implementación de los principales productos de analítica en nuestras entidades.

3.1.4.1. Definir el modelo de datos

Para organizar los datos de una forma estructurada que permita hacer consultas y análisis, es necesario:

- **Alinear el modelo de datos con las necesidades:** construir una estructura de datos (tabla(s), hoja(s) de cálculo, etc.) que permita consultar los datos relacionados de forma fácil y ágil a partir de la necesidad de datos e información (a través de las preguntas o procesos a impactar).
- **Definir el tipo de repositorio para los datos** según los tipos de datos a almacenar, por ejemplo, para datos estructurados que cambian frecuentemente se recomienda una base de datos o, en caso de requerirse consultas históricas de grandes volúmenes de datos, se recomienda una bodega o lago de datos ([ver anexo A. Ejemplo de arquitectura del flujo de datos para la analítica institucional](#)).
- **Identificar las fuentes de datos disponibles** para el llenado del repositorio y modelo de datos definidos ([ver anexo A. Ejemplo de arquitectura del flujo de datos para la analítica institucional](#)).
- **Construir los procesos** que permitan que los datos de las fuentes disponibles pasen (se almacenen) de forma periódica, y en lo posible automática, en el repositorio o modelo de datos definido ([ver anexo A. Ejemplo de arquitectura del flujo de datos para la analítica institucional](#)) cumpliendo con los lineamientos de gobierno de datos definidos por la entidad.

Adicionalmente, es necesario validar:

- Que los datos se almacenen en el modelo de datos definido (tablas, hojas de cálculo, entre otros).
- Que no existan otras estructuras de datos distintas al modelo de datos planteado.
- Que el modelo definido permita responder a la necesidad planteada por medio de consultas.

3.1.4.2. Definir los indicadores clave de desempeño (KPI)

Para que un indicador sea considerado KPI este debe ser SMART (por la sigla en inglés de Specific, Measurable, Attainable, Relevant, Time-bound), es decir, que sea específico y medible, asociado a objetivos claros, que dé información del progreso y desviación hacia la meta, alcanzable y relevante (que genere valor) en un marco de tiempo adecuado. Por ejemplo, Kerzner (2017) plantea algunos indicadores que pueden servir de referente para la medición en nuestras entidades:

- **Retorno de la inversión (ROI, por la sigla en inglés de *Return On Investment*):** permite entender si un proyecto, acción realizada o adquisición de un producto o intangible, realmente genera un retorno económico por encima de la inversión realizada. Por ejemplo, si el desarrollo de un sistema automático de caracterización de la ciudadanía genera ahorros superiores al costo de implementación del sistema.
- **Valor presente neto:** permite calcular el valor a hoy de las inversiones, costos o retornos futuros de un proyecto, *software*, etc. por ejemplo, si la entidad debe invertir un valor mensual durante dos años por concepto de implementación, soporte y mantenimiento de un sistema de caracterización de la ciudadanía, en el presupesto de esos dos años debemos

tener en cuenta la inflación para calcular cuánto dinero correspondería si tuviéramos que desembolsarlo hoy.

- **Periodo de recuperación:** permite cuantificar el tiempo mínimo necesario para llegar a un saldo cero a partir de una inversión realizada, es decir, cuánto tiempo se requiere para recuperar la inversión. Para nuestro ejemplo de implementación de un sistema de caracterización de la ciudadanía, corresponde con el tiempo en el cual el retorno (ahorro en costo de digitación manual de los datos) y la inversión en el sistema de caracterización de la ciudadanía (implementación) llegan a cero.
- **Reducción de costo:** permite cuantificar el porcentaje de reducción en los costos a partir de la intervención realizada. Para nuestro ejemplo de implementación de un sistema de caracterización de la ciudadanía correspondería a:

$$\text{PORCENTAJE (\%) DE AHORRO EN COSTOS} = \frac{\left[\begin{array}{l} \text{costo caracterización} \\ \text{y digitación manual} \\ \text{de los datos} \end{array} - \begin{array}{l} \text{costo de caracterización} \\ \text{empleando el sistema} \\ \text{de caracterización de la} \\ \text{ciudadanía} \end{array} \right]}{\text{costo de caracterización y digitación manual de los datos}} \times 100 \%$$

Si el porcentaje es menor que 0 % (valores negativos) significaría que es más costoso caracterizar un ciudadano con el sistema de caracterización de la ciudadanía que hacerlo de forma manual (como se hacía antes), en caso de que sea mayor a 0% significará que es más económico emplear el sistema de caracterización de la ciudadanía.

- **Reducción de trámites:** permite cuantificar el porcentaje de reducción de los trámites realizados por los usuarios a partir de la intervención realizada. Se puede complementar con indicadores de aumento en la satisfacción por el usuario/ciudadano en el trámite realizado o el porcentaje de respuesta en primer contacto (el ciudadano pudo completar el trámite en una única atención por parte de la entidad).
- **Número de nuevos clientes:** en nuestro contexto correspondería al número de nuevos ciudadanos atendidos, lo que permite conocer cuántos ciudadanos nuevos (que no hayan realizado el trámite en el pasado) realizaron el trámite o fueron atendidos en un proceso. Se puede complementar con indicadores: porcentaje de ciudadanos que realizan el trámite continuamente (que lo realiza al menos una vez al mes de forma consecutiva) o porcentaje de ciudadanos que realizan el trámite de manera virtual.
- **¿El objetivo del indicador es relevante?:** el indicador debe dar información de valor para tomar alguna decisión alineada con la misión de la entidad; en caso de que no, este indicador debe ser replanteado.

En caso de que no existan los indicadores requeridos para el análisis de datos planteado, es necesario definir nuevos indicadores al interior de nuestra entidad, para lo cual se recomienda:

- **Definir claramente el objetivo del indicador:** el objetivo de medición del indicador debe ser lo más preciso posible y estar alineado con la misión de la entidad, podemos preguntarnos: ¿el objetivo del indicador es relevante? ¿qué pasaría si no existiera este indicador, existirá alguno que lo reemplace?
- **Validar que el indicador pueda ser medido:** la entidad debe tener una fórmula matemática que le permita calcular con los datos existentes el valor del indicador, de modo que puedan saber si se cumplió o no con el objetivo planeado.

- **Validar con las áreas y grupos de valor que lo utilizarán:** es necesario que los servidores que interactuarán con el indicador conozcan cuál es su objetivo, cómo se calcula y, además, cuenten con el conocimiento necesario para su interpretación.
- **Definir los valores de referencia para el indicador:** para validar si se está cumpliendo con el objetivo asociado al indicador, es necesario contrastar el valor del indicador con algún otro valor de referencia, por ejemplo, una meta a alcanzar en la que el porcentaje de cumplimiento del presupuesto sea del 95 % o un valor mínimo a cumplir en el que el porcentaje de satisfacción del ciudadano ante la atención en ventanilla sea de, al menos, un 80 %.
- **Gestionar y monitorear el indicador:** validar periódicamente el valor del indicador (diariamente, semanalmente, mensualmente) comparándolo con el valor de referencia definido, en caso de que difiera (no se cumple la meta o está por debajo del valor mínimo definido) la entidad debe revisar la(s) razón(es) y tomar las acciones necesarias para el cumplimiento de la meta o mínimo definido. Para lograr lo anterior se recomienda asignar a un responsable o líder del indicador que realice estas tareas.
- **Gestionar el uso del indicador:** validar que las personas que requieren usar el indicador puedan hacerlo, esto incluye que: comprendan cuál es el objetivo y cuál es la forma de cálculo, que puedan consultarlo cuando lo requieran y que exista algún mecanismo para resolver dudas o cambiarlo si es necesario.
- **Validar y actualizar el indicador a partir de cambios en los datos empleados para el cálculo:** es necesario validar de forma periódica (trimestral, semestral, anual) si los datos empleados para el cálculo del indicador han sufrido algún cambio en su definición o almacenamiento, para ello debemos revisar si el cálculo del indicador sigue siendo

coherente, en caso contrario, es necesario actualizarlo y comunicar la novedad a las personas que lo utilizan.

3.1.4.3. Implementar procesos automáticos de cálculo de indicadores

Para que los indicadores estén actualizados es necesario que puedan calcularse de manera periódica, si se hace de forma manual implicaría un esfuerzo por parte de las personas encargadas en la entidad, lo que limitaría su capacidad para realizar actividades de análisis. Por lo anterior, se recomienda la implementación de procesos automáticos de cálculo de indicadores, es decir, sin intervención humana, para ello, consideremos:

- Definir claramente los datos requeridos y dónde encontrarlos (fuente de los datos).
- Definir la periodicidad de cálculo y actualización del indicador, de acuerdo con los datos requeridos, por ejemplo, si los datos se actualizan de forma mensual, no sería necesario actualizar el indicador de forma diaria.
- Cumplir con todos los lineamientos del gobierno de datos definidos por la entidad, entre ellos, utilizar las herramientas de procesamiento de datos y repositorios de información dispuestos para este fin.
- Definir cuáles cálculos anteceden y preceden a los demás para optimizar el tiempo de cálculo.
- Generar visualizaciones o informes que permitan identificar cuáles indicadores han sido actualizados, cuáles no y por qué.
- Definir al menos un responsable de la gestión y monitoreo del proceso de actualización del indicador para identificar problemas en el cálculo (valores incoherentes, fallo en el cálculo o al almacenar el indicador, entre otros).

3.1.4.4. Elaborar reportes y tableros de control

Para Sherman (2015), los tableros de control o *dashboards* son una herramienta que permite visualizar información numérica y gráfica de distintas fuentes en una única vista, cuyo objetivo es contar historias con datos de forma fácil y ágil.

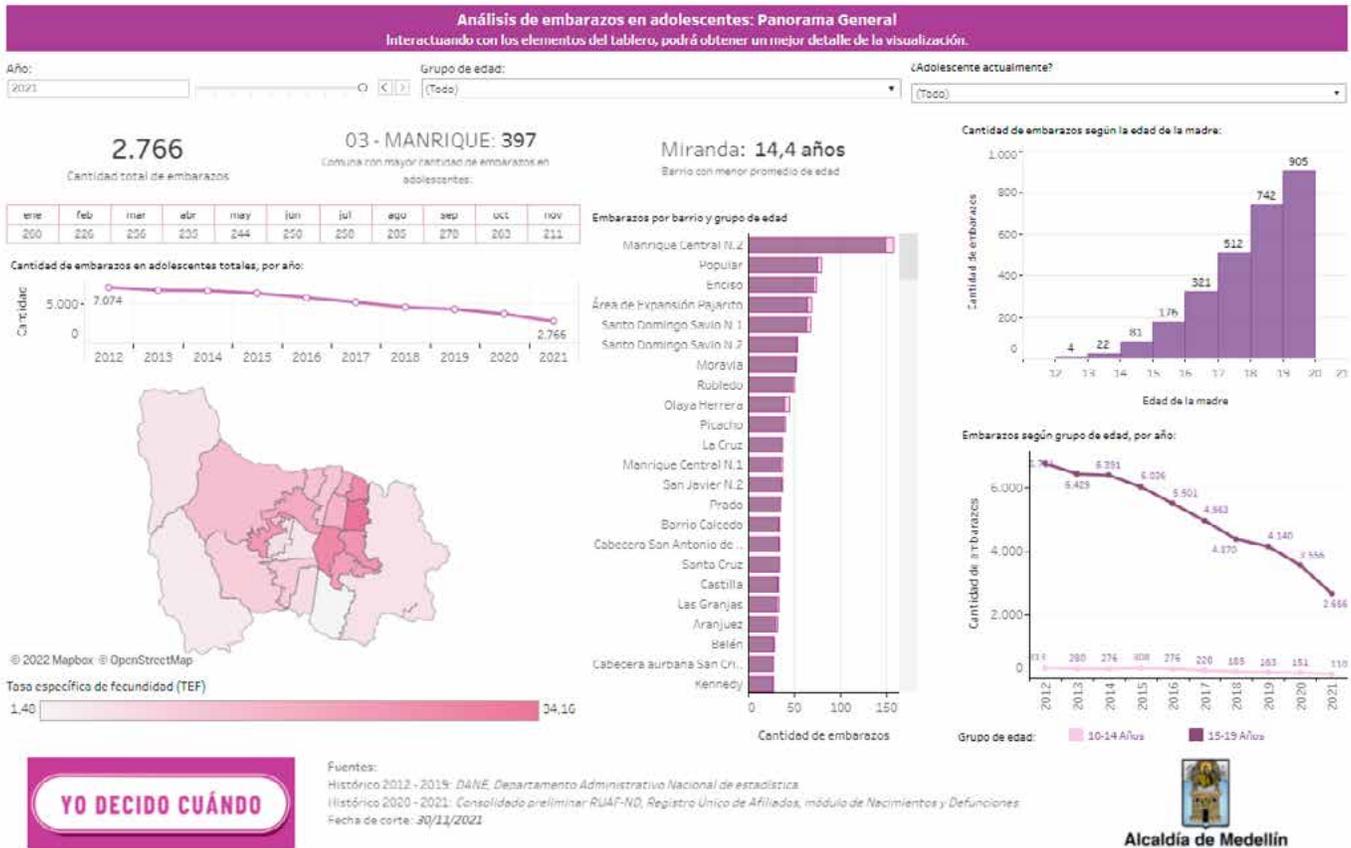
Al momento de visualizar datos en tableros de control o *dashboards* y reportes se recomienda:

- **Identificar el propósito del tablero o *dashboard***, es decir, definir claramente el para qué del tablero de control o *dashboard* y cuales elementos, datos o indicadores son realmente relevantes.
- **Escribir y validar la historia que se va a contar con los datos – el mensaje es el “rey”**, es decir, identificar cual será la secuencia de análisis recomendada al usuario del tablero de control, cuales datos o indicadores serán presentados primeros y cuáles de últimos, para ello existen dos formas:
 - ▶ Desde la conclusión (indicador/dato principal) explicando paso a paso a través de los elementos visuales del tablero de control o *dashboard* cómo llegó a ese dato o indicador. Por ejemplo, en la Figura 7 se muestran los principales indicadores y luego se explica estos totales a partir del comportamiento por municipio.
 - ▶ Desde el detalle (indicador/dato desagregado) explicando cómo agruparon los datos y qué ocurrió con ellos hasta llegar a la conclusión (indicador/dato principal). Por ejemplo, en la Figura 8, en la que se inicia analizando el comportamiento del PIB, su variación y distribución por actividad, concluyendo con el respectivo porcentaje de variación nacional y por departamento.

Ambos son válidos, lo que debemos tener en cuenta es la coherencia del discurso a través de los elementos visuales.

- **Entender el público a quien va dirigido el tablero o *dashboard***, si es un público técnico se deberán usar indicadores técnicos que sean necesarios para sustentar la conclusión. Por el contrario, si el público no es técnico, el mensaje deberá mostrar en un lenguaje sencillo cómo se llegó a la conclusión.
- **Si no funciona a lápiz y papel no va a funcionar en un tablero de control**, es decir, es necesario realizar un boceto de la visualización que nos permita entender cómo se vería, qué partes tendría y qué elementos se utilizarían con el fin de validar y realizar las correcciones conceptuales necesarias para lograr el objetivo de la visualización.
- **No saturar – menos, es más**, es decir, los tableros de control deben utilizar la menor cantidad de elementos visuales para lograr su propósito.
- **Los tableros de control no deben centrarse en un dato específico, sino en la explicación de los datos (tendencia, comparación, crecimientos y decrecimientos, entre otros)**, permitiendo interactuar con sus elementos visuales y entender los datos y su comportamiento mediante la evolución del indicador en la entidad. Por ejemplo, en la Figura 7 se muestra el total y la tendencia de los embarazos en el municipio de Medellín por año y grupo de edad, lo que permite entender su evolución y disminución en el tiempo.

Figura 7. Análisis de embarazos adolescentes (panorama general) para el municipio de Medellín



Nota. Tomado por Función Pública (2022) del portal de datos abiertos de Medellín, MEDATA, 2021.

- **Usar la imagen institucional en todas las visualizaciones**, cumpliendo con los lineamientos definidos en el manual de imagen de la entidad.
- **Usar un esquema autocontenido**, es decir, que las visualizaciones, los reportes, informes y tableros de control no deben necesitar de alguien para entenderlas, deben tener sentido por sí solas, para lo cual se puede utilizar elementos visuales de uso común, por ejemplo, los que se describen en el Anexo C. Tipos de visualizaciones de datos.

3.1.4.5. Contar con sistemas o programas de servicio de consulta de datos específicos

Para consultar, seleccionar y filtrar datos específicos sin la necesidad de requerimientos a otras áreas, es necesario que la entidad cuente con sistemas o programas de *software*. Para diseñarlos o usarlos debemos tener en cuenta:

- Que sean fáciles de usar y que no requieran un conocimiento avanzado en tecnología.
- Que permitan filtrar los datos de forma automática de acuerdo con el rol y funciones del servidor(a), es decir, que cumpla con el gobierno de datos de la entidad.
- Que permitan cruzar datos de distintos repositorios y que estos puedan ser visualizados fácilmente.
- Que permitan seleccionar aquellos datos que se requieran sin necesidad de programación.
- Que cuenten con manuales y guías acerca del uso del sistema o programa de *software*.
- Que permitan validar que los datos generados correspondan con aquellos almacenados y que no existan transformaciones adicionales que no hayan sido definidas y socializadas con los usuarios.

3.1.4.6. Crear o usar modelos matemáticos

Para la creación o uso de modelos matemáticos recomendamos:

- **Definir la metodología a emplear**, esto le permitirá responder de forma adecuada a la necesidad que se pretende resolver (procesos, planes, programas, proyectos, etc.). En el caso de implementar analítica predictiva se suele utilizar la metodología CRISP-DM (por la sigla en inglés de *CRoss-*

Industry Standard Process for Data Mining) (ver Anexo B. Minería de datos y *big data*).

- **Definir la técnica o algoritmo a emplear** a partir de la necesidad que se pretende resolver y los datos disponibles.
- **Incluir el conocimiento generado** a partir de los datos en el proceso de modelación, por ejemplo: tendencia, ciclo, estacionalidad, correlación y causalidad.
- **Plantear varios modelos con parámetros distintos** que permitan identificar si la técnica o algoritmo es apropiada o los resultados obtenidos son coincidentes.
- **Validar los resultados del modelo** a través de una métrica acorde con la técnica o algoritmo empleado.
- **Evaluar posibles sesgos en los datos** que predispongan los modelos matemáticos a resultados parciales.
- **Validar que el modelo no esté memorizando los datos**, es decir, que incorpore el comportamiento de los datos y no arroje un valor incoherente ante nuevos datos.
- **Para los modelos de optimización, definir criterios (*drivers*) de optimización apropiados** de acuerdo con la(s) pregunta(s) a resolver y sus restricciones, por ejemplo, si debemos definir el programa de alimentación que minimice el tiempo en que los niños con malnutrición recobren el estado de nutrición adecuado, el primer criterio debería ser: que sea más probable o efectivo el volver al estado de nutrición normal; sin embargo, si existen limitaciones económicas se puede incluir: al menor costo posible; si existen restricciones en tiempo: que pueda realizarse en menos de x meses; y si existen restricciones geográficas o de desplazamiento, podríamos incluir que: la sede de atención este a menos de 1 km del hogar de nuestros niños.
- **Sugerir más no obligar a seleccionar la respuesta óptima** cuando en los datos puedan existir sesgos o falte

información relevante para tomar la decisión. Por ejemplo, si nuestro municipio se ha caracterizado por contar con población menor de 35 años y hemos realizado modelos matemáticos que seleccionan aquellos más vulnerables para la inclusión en programas sociales, pero en el último mes se han mudado al municipio una gran cantidad de adultos mayores de 70 años, posiblemente no se tenga información histórica que permita determinar adecuadamente su grado de vulnerabilidad, en este caso, se debe sugerir el grado de vulnerabilidad y el programa a ser incluido, y proponer su evaluación por algún servidor(a) para que la participación en los programas no este sesgado por los datos.

3.1.5. Validar y evaluar los resultados obtenidos

Una vez la entidad implemente los productos de analítica, es necesario:

- Validar el uso y apropiación de los productos de analítica por parte de los servidores de la entidad, para ello pueden revisar si la entidad usa los resultados de los análisis en los planes, programas, proyectos, etc. que desarrolla, también puede llevar a cabo encuestas periódicas o entrevistas con personas usuarias del modelo, entre otros.
- Validar si el producto de analítica le ha generado valor a la entidad en términos de eficiencia, eficacia, generación de ahorros y bienestar a los grupos de valor, entre otros.
- Validar que los productos continúen siendo coherentes con los datos de la entidad, para ello debe revisar periódicamente si la medida de validación de desempeño sigue estando en los mínimos o máximos definidos, en caso contrario, actualice el modelo matemático.
- En caso de que el producto de analítica ya no sea necesario o útil para la entidad deben eliminarlo.

Conclusiones

- Los datos e información sin ser usados representan un costo para nuestras entidades, dado que almacenarlos y hacer visualizaciones (presentaciones, tableros de control e informes) requiere el uso de tecnología (computadores, repositorios de datos o *software*, entre otros) y el esfuerzo de personas.
- La analítica institucional nos permite ser más efectivos, precisos y eficientes, disminuir los costos, errores y riesgos e identificar oportunidades no evidentes.
- Todos los tipos de analítica son importantes y necesarios para la toma de decisiones en nuestras entidades, cada uno permite dar respuesta a un tipo de pregunta distinta.
- El talento humano es el pilar más importante para la implementación de la analítica institucional en nuestras entidades, por lo tanto, debe comprender qué es la analítica y cómo aplicarla a los procesos de la entidad.
- Los productos de analítica institucional son herramientas para la gestión del conocimiento en las entidades públicas.
- Los resultados de la analítica de datos no dan soluciones por sí solos, de modo que es necesario que estos repercutan en la realidad de la entidad (planes, programas, proyectos y necesidades, entre otros) y no queden en un ejercicio meramente académico.

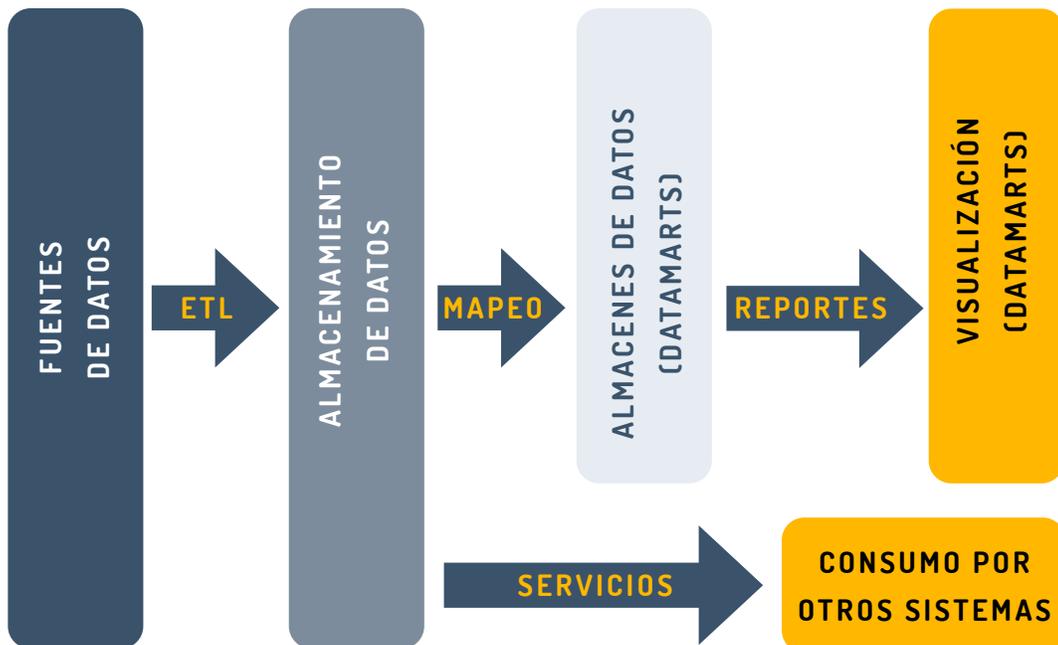


Anexos

Anexo A. Ejemplo de arquitectura del flujo de datos para la analítica institucional

Para llevar a cabo la analítica institucional es necesario adquirir, procesar y presentar los datos de acuerdo con las necesidades y preguntas planteadas por la entidad. Para ello, podemos plantear una arquitectura de referencia para el flujo de datos en el proceso de análisis, en la que se incluya, entre otros: la adquisición de datos, el almacenamiento, el mapeo con conceptos del área (equivalencia entre los datos almacenados y su representación o uso) y su posterior visualización y consumo en otros sistemas (ver Figura 1).

Figura 1. Ejemplo de arquitectura del flujo de datos



Nota. Adaptado por Función Pública (2022) con base en el documento *Data Warehousing Fundamentals for IT Professionals* de Paulraj Ponniah (2011), p. 35.

A continuación, desarrollaremos los componentes del modelo de arquitectura para el flujo de datos propuesto en la Figura 1:

¿Cuáles son las fuentes de datos?

Según Ponniah (2011), las fuentes de datos hacen referencia a sistemas, bases de datos, archivos y similares, que se pueden agrupar así:

- **Sistemas de gestión:** en estos encontramos los planeadores de recursos empresariales (ERP, por la sigla en inglés de *Enterprise Resource Planning*), los gestores de procesos de negocio (BPM, por la sigla en inglés de *Business Process Management*), manejadores de la relación con el cliente (CRM, por la sigla en inglés de *Customer Relationship Management*) y otros, que tienen como objetivo la administración de procesos al interior de nuestra entidad, tales como: contables, propios (*core*), de relación con el cliente, marketing, servicio, operación, etc.
- **Sistemas transaccionales:** corresponden a la información generada por todos aquellos sistemas procesan y administran transacciones, por ejemplo: sistema de atención al usuario, sistema de caracterización del ciudadano o sistema de gestión de turnos, entre otros. Entre las principales fuentes encontramos las bases de datos relacionales y no relacionales que soportan los procesos misionales de nuestras entidades.
- **Bodegas de datos (DWH, por la sigla en inglés de *data warehouse*):** de acuerdo con Ponniah (2011), corresponde a "... *The data warehouse is an informational environment that: Provides an integrated and total view of the enterprise. Makes the enterprise's current and historical information easily available for strategic decision making ...*" (p. 15), es decir, son un sistema de almacenamiento que proporciona una visión integral de la entidad con el objetivo de analizar

información actual e histórica de forma fácil para la toma de decisiones.

- **Archivos** que tenemos al interior de nuestros procesos, normalmente incluyen: archivos de texto plano (sin formato), hojas de cálculo, archivos en formato propietario o cerrados (definidos por los sistemas que lo generan).
- **Flujos de datos:** incluyen aquellos que son entregados de forma ordenada y constante en el tiempo correspondientes a sensores (internet de las cosas, vestibles, entre otros), web a través de servicios: API, sistemas internos o plataformas especializadas de mensajería, entre otros.

Es importante identificar y gestionar los tipos de fuentes de datos conforme al gobierno de datos que haya definido nuestra entidad.

Procesos de ETL (*Extract-Transform-Load*) de datos

Según Sherman, la extracción-transformación-carga (ETL, por la sigla en inglés de *Extract-Transform-Load*) es un proceso por medio del cual se extraen los datos desde una fuente de datos específica, para luego transformarlos y cargarlos en un destino (2014, p. 17). La Figura 2 muestra el proceso de ETL:

Figura 2. Proceso de ETL



Nota. Elaborado por Función Pública (2022) con base en Sherman, R. (2014) pp. 76-78 y Kimball, R. & Ross, M. (2013) p. 422.

A continuación, desarrollaremos los elementos que componen la Figura 2 sobre el proceso de ETL (extracción-transformación-carga) de datos:

Extraer

En este proceso se utiliza el archivo o tabla de mapeo para consultar y extraer los datos. Para ello, se emplean programas especializados de ETL, programas de uso general adecuados para este proceso implementados en nuestra entidad o consultas SQL que se exportan periódicamente. Estos obtendrán los datos de la fuente y los llevarán a un espacio de almacenamiento temporal para su posterior procesamiento.

Cuando se emplean programas especializados de extracción (incluye aquellos de replicación), es necesario configurar los accesos (credenciales, usuarios, tipo de fuente de datos, tablas/hojas, campos/atributos, tipo de extracción), los mecanismos de extracción de datos (consulta, replicación, otro) y la periodicidad. En el caso de implementación propia, se debe considerar el mecanismo de ejecución periódica que se va a utilizar, además de los recursos tecnológicos necesarios.

Según Kimball y Ross (2013), los criterios para identificar el tipo de extracción por realizar pueden ser:

- **Completa (*full*):** todos los datos de la estructura (tabla, hoja, archivo) sin restricción o validación por alguna característica.
- **Incremental/delta:** los datos que no hayan sido extraídos previamente se consultan a partir de alguna característica de la estructura (tiempo, identificación, etc.). Este tipo de extracción es un subconjunto menor o igual a la completa (es igual únicamente en la primera extracción).

Transformar

La transformación de datos corresponde a cambiar el dato por algún cálculo o reemplazo a partir de unas reglas definidas. Puede suceder durante el proceso de extracción (en la memoria temporal del programa empleado) o después de la carga en el espacio temporal de almacenamiento, denominado zona de aterrizaje (*staging*), que corresponde a un espacio dentro del destino de los datos (suele ser una base de datos o similar) y se utiliza para realizar las operaciones de transformación previas a la carga en el modelo final (destino). Si el destino es un archivo, estas operaciones se suelen realizar durante el proceso de extracción y transformación.

Para lograr la transformación se utilizan rutinas en SQL (archivos de código en lenguaje SQL, que puede ser ejecutado en un manejador de base de datos o bodega de datos para obtener datos), programas especializados de ETL (programas que han sido diseñados e implementados para la ejecución de los pasos de extracción-transformación-carga de forma eficiente) y programas de uso general (programas que pueden emplearse para realizar distintas tareas, desde cálculos sencillos hasta algoritmos de procesamiento de datos) adecuados para este proceso.

Cargar

La carga corresponde a llevar los datos ya transformados a su repositorio o modelo de datos final. En el caso de bases de datos o similares, corresponde a almacenar los datos en ese modelo de datos dispuesto para ello; cuando son archivos, corresponde a crear el archivo con los datos generados.

Es de anotar que cuando se realiza el proceso de ETL en programas especializados de ETL, el área de almacenamiento temporal corresponde a la memoria propia (física-lógica) de dicho programa, haciendo las veces de extracción y transformación de manera interna por medio de una malla de procesamiento.

Un ejemplo en nuestras entidades puede ser el proceso de paso de información desde los sistemas de caracterización de ciudadanos y el de entrega de beneficios al ciudadano, a un único sistema de almacenamiento (base de datos, hoja de cálculo, repositorio, bodega/almacén de datos) que permita integrar los datos y conocer si realmente estamos cubriendo institucionalmente a la ciudadanía o grupos de valor.

Almacenamiento de datos

Ponniah (2011) plantea que uno de los objetivos del almacenamiento es contar con un repositorio transversal de datos organizados que, en nuestro caso, le permita a la entidad pública acceder a estos de forma ágil, rápida y eficiente. Para ello, es necesario que los datos estén integrados y disponibles en una bodega de datos (DWH) de forma física o lógica, aprovechando las bases de datos relacionales y no relacionales, los lagos de datos y similares.

A continuación, enunciaremos las principales características de los modelos de almacenamiento, a partir de lo propuesto por Ponniah (2011):

Bodega de datos física (DWH)

- Sistema estructurado (como los de las bases de datos o hojas de cálculo) con un modelo propio diseñado para un propósito específico, por ejemplo, información pública, en el cual las aplicaciones y la infraestructura están fuertemente cohesionadas, es decir, no es posible instalar la aplicación en otra infraestructura.
- Se posiciona como un unificador de datos transversal con procesos de extracción-transformación-carga (ETL, por su sigla en inglés) en una malla de procesamiento que se ejecuta periódicamente para la actualización de los datos.
- Cuenta con infraestructura propia, tanto local o en sitio (*on premise*) como en nube (*cloud*).

- Está enfocada en representar la entidad a partir de tablas y vistas con modelos de indicadores transversales y propios por área.
- Las vistas suelen ser usadas como capa de seguridad (permite darle acceso a cierto contenido de las tablas a los usuarios por medio de consultas-vistas) y para la optimización de consultas (precargando los datos en memoria para su uso posterior).
- Corresponden a sistemas de análisis de datos que pueden funcionar asíncronamente, y, por lo tanto, no son considerados sistemas transaccionales.

Bodega de datos lógica

Bodega de datos que adopta las características de un DWH, pero empleando infraestructura compartida, sistemas de almacenamiento alternos y representaciones lógicas de ellas. Entre sus características principales se encuentran:

- Modelo híbrido, basado principalmente en un lago de datos (*data lake*) y un manejador de base de datos relacional o No SQL.
- Implementa modelos de datos enfocados a bodegas de datos, los más representativos son aquellos basados en estrella y copo de nieve.
- La infraestructura no es dedicada, es decir, que los equipos no se emplean únicamente para la bodega de datos, y pueden alojarse en nubes públicas, privadas o híbridas.
- Existe correspondencia entre tablas y vistas para la consulta de datos.

Para definir el mejor esquema de almacenamiento que equilibre la cantidad de datos almacenados y los tiempos de respuesta, es necesario analizar el nivel de concurrencia de usuarios (cantidad de

usuarios que realizan una operación al mismo tiempo) y el nivel de consulta de los datos (cantidad en un periodo de tiempo).

Por otra parte, según Kimball y Ross (2013), los datos se guardan en una tabla equivalente a una hoja de cálculo que relaciona los campos y atributos en las columnas y los registros en las filas.

Mientras que, según Microsoft (2021), "... una vista corresponde a una tabla virtual cuyo contenido está definido por una consulta ...", es decir, no contiene los datos en un espacio dentro de la bodega de datos y son generados al momento de consultar, por ejemplo, se puede contar con una tabla "Ciudadano" que tiene las fechas de nacimiento de toda la ciudadanía, sin embargo, se puede crear una vista "VCiudadano" que contenga la edad del ciudadano al día de hoy.

El uso de vista tiene ventajas como:

- Combinar columnas/filas desde múltiples tablas u otra vista y tener una vista consolidada.
- Generar vistas con base en el acceso del usuario (que se determinan según el rol, cargo, funciones al interior de la entidad y temas a su cargo, entre otros), sin que tenga acceso directo a la(s) tabla(s).
- Actuar como una capa abstracta para los sistemas que emplean los datos, por lo que cualquier cambio en las tablas de la base de datos (estructura y tipo de dato, entre otros) sin que genere un reproceso en dichos sistemas debido a que consultan una vista con la misma estructura.

Sin embargo, la principal desventaja del uso de vistas radica en que, al no estar almacenada en la base de datos o bodega de datos física, sus datos se generan al momento de consultarla, por lo tanto, puede tener un tiempo de respuesta alto. Para evitar lo anterior, existen las vistas materializadas que sí almacenan los datos de forma periódica en la base de datos o almacén de datos física.

Por último, los modelos de almacenamiento tipo lago de datos (DL, por la sigla en inglés de *datalake*), son un tipo de almacenamiento que permite crear y gestionar el equivalente a tablas y vistas, este tema será tratado en el Anexo B. Minería de datos y *big data*.

Mapeo

Según Kimball y Ross (2013), el mapeo se refiere al proceso de identificar la información de una fuente de datos y relacionarla con una de destino (como se puede apreciar en la Figura 3).

Figura 3. Estructura conceptual del proceso de mapeo



Nota. Elaborado por Función Pública (2022) con base en Kimball, R. & Ross, M., (2013) p. 529.

Como se muestra en la Figura 3, para mapear los datos es necesario:

- Identificar para cada **fuentes** de datos su estructura (campo/ atributo), lo que incluye: nombre, tipo de dato (cómo se almacena en la fuente), longitud o tamaño, descripción o significado del campo.
- Identificar cuál va a ser el campo de **destino** donde se va a guardar el dato de la fuente, este debe incluir: nombre, tipo

de dato, longitud, transformaciones (descripción y forma de cálculo), descripción o significado del campo.

Una vez ha mapeado la fuente y el destino, se debe registrar la información identificada en un archivo (hoja de cálculo) o **tabla de equivalencias** (bases de datos) accesible para las distintas áreas y equipos que los requieran.

Almacenes de datos (*datamarts*)

Según Sherman (2014), los almacenes de datos o *datamarts* pueden entenderse como *“A subset of a data warehouse that’s usually oriented to a business group or process”* (p. 22), es decir, son un subconjunto de la bodega de datos que puede estar orientado a un tema o proceso específico.

Se pueden identificar dos tipos principales de estructura presentes en las arquitecturas de los almacenes de datos:

OLAP (*On-Line Analytical Processing*)

“Los sistemas OLAP son bases de datos orientadas al procesamiento analítico” (Sinergia, s.f.) y son los más empleados por los almacenes de datos. Este análisis implica la lectura de grandes cantidades de datos para extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos, etc.

OLTP (*On-Line transactional Processing*)

“Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones” (Sinergia, s.f.). Una transacción suele involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales.

La entidad puede o no utilizar almacenes de datos en la arquitectura de análisis de datos definida; sin embargo, estos facilitan la consulta y análisis de datos. En caso de implementar tableros de control con una plataforma de BI, esta generará automáticamente las estructuras necesarias para el correcto procesamiento y consulta de la información en los informes y tableros, las cuales corresponden con almacenes de datos.

Visualización

La visualización es uno de los principales medios para mostrar los resultados del proceso de integración (pasos ETL y almacenamiento) y análisis de los datos realizado por las personas de la entidad. Según Wexler, *et al.* (2017, p.16), algunos de los tipos de visualización más empleados son:

- **Reportes:** informes en una plataforma de análisis de datos que se actualizan de forma automática, pero no permiten interacción.
- **Tableros de control o dashboard:** reportes interactivos (tableros) que contienen objetos visuales (gráficos, texto, tablas, mapas, entre otros) que muestran datos y permiten interactuar con ellos.
- **Presentaciones:** diapositivas donde se presentan los resultados de un análisis de datos, suelen ser estáticos, aunque en ocasiones es posible conectarlo a la fuente de datos para su actualización automática.

Sherman (2014) hace una distinción entre operacionales y analíticos, los cuales podemos entenderlos como:

Operacionales: son aquellos que dan respuesta del ahora y están asociados principalmente a la gestión de procesos. Algunas de sus características son:

- Fuente de datos transaccional.
- Generan o calculan un indicador clave para continuar con el proceso.
- Garantizan tiempos de respuesta casi en línea.
- Las consultas se realizan directamente al sistema transaccional (no admite procesos *batch*).
- Generan puntos de decisión en el proceso.

Analíticos: son aquellos enfocados a un análisis histórico y de tendencia para identificar cómo vamos. Algunas de sus características son:

- Responden a indicadores de relevancia para la entidad (operativos y estratégicos).
- La fuente es una bodega (DWH) o un almacén de datos (*datamart*).
- Pueden estar pre ejecutados (*batch*) con tiempos de respuesta adecuados a las necesidades de presentación.
- Suelen asociarse a decisiones transversales (p. 103).

Para implementar los tableros debemos considerar si son de tipo transaccional o analítico, dado que de esto depende qué datos se consultan en línea (consultas directas a la fuente de datos), o a través de un almacenamiento de una memoria tipo caché (pre ejecutados).

En el día a día de nuestras entidades, un ejemplo de tablero operacional podría ser aquel enfocado en el proceso de atención a la primera infancia, en el cual podemos identificar cuántos niños han sido incluidos en los programas de alimentación disponibles, cuántos refrigerios o similares se les han brindado a ellos, entre otros. Mientras que un tablero analítico alrededor del mismo tema podría ser el análisis de la evolución del estado de nutrición de nuestros niños, identificando las causas y planteando posibles escenarios de intervención.

Consumo por otros sistemas

Como se ve en la Figura 1, para el proceso de análisis de datos se pueden usar los datos y visualizaciones en otros sistemas través de servicios web y API's. Para lo cual, es necesario identificar los datos requeridos por el otro sistema, cuál es la periodicidad de envío o consumo de información, cuál es el método de integración, y cuál es el esquema de seguridad y validación de accesos. Por ejemplo: si solo se requiere analizar datos a través de tableros de control, posiblemente no se incluyan mecanismos de consumo por otros sistemas.

Anexo B. Minería de datos y *big data* (datos masivos)

En este anexo nos enfocaremos en identificar qué es la minería de datos y cómo se implementa a través de la metodología CRISP-DM (por la sigla en inglés de *C*Ross-*I*ndustry *S*tandard *P*rocess for *D*ata *M*ining). Adicionalmente, describiremos las características propias del *big data*.

Minería de datos

Según Jones (2019), "... la minería de datos se refiere al proceso de depuración, recopilación, procesamiento, análisis y extracción de información relevante de datos ..." (p. 1). Es decir, es tomar los datos, procesarlos y modelarlos al servicio de la identificación de patrones, como se puede apreciar en la Figura 1.

Figura 1. Representación del proceso de minería de datos



Nota. Adaptado por Función Pública (2022) con base en Jones (2019) p. 2.

Para implementar la minería de datos es necesario usar una metodología, que para este caso sugerimos emplear la metodología CRISP-DM. Para usar esta herramienta Pérez (2022, pp. 1-12) plantea seis pasos que le ayudarán a llevar a cabo la minería de datos en su entidad. A continuación encontrará una síntesis del paso a paso:

Entender el objetivo del área: se enfoca en la comprensión de los objetivos y exigencias del proyecto de minería de datos o pregunta analítica a resolver desde una perspectiva de área.

Entender los datos: cubre la recolección de datos inicial y las actividades que permiten describir los datos, identificar su calidad y descubrir conocimiento preliminar a partir de ellos.

Preparar los datos: cubre todas las actividades necesarias para construir el conjunto de datos final para iniciar la modelación matemática.

Construir el modelo (matemático): se aplican las distintas técnicas de minería de datos para dar respuesta a la pregunta planteada.

Evaluar el modelo: en este se evalúa el modelo construido en el paso 4 para determinar si es útil a las necesidades de la entidad, para ello pueden partir de criterios de error basados en distancia (diferencias cuadráticas/absolutas entre el valor generado y el valor real, entre otros), variabilidad (medidas de dispersión de los datos, diferencia de varianzas, entre otros) u otro.

Despliegue del modelo: son todas las actividades necesarias para que sea posible usar los modelos al interior de la entidad.

Para hacer un abordaje ordenado y alineado a la solución de la necesidad planteada, se recomienda contar con una base de conocimiento que permita replicar modelos realizados entre las distintas áreas y dependencias de nuestra entidad.

Big data (datos masivos)

De acuerdo con el Departamento Nacional de Planeación (2021), las características del *big data* son: velocidad, variedad, veracidad, volumen y valor, como se muestra en la Figura 2:

Figura 2. Características del big data



Fuente: Imagen tomada del documento Aprovechamiento de datos para la toma de decisiones en el sector público del Departamento Nacional de Planeación (2021), p. 22.

Para almacenar, procesar y consultar datos que cumplan con las características del *big data* existen nuevas tecnologías como los lagos de datos que permiten el almacenamiento en sistemas de archivos más allá que estructuras de tablas. Lo anterior implica que los procesos ETL (extracción-transformación-carga) cambien por mecanismos de ingesta de datos (copia de los datos en crudo) y replicación de datos (copia de datos directamente de la fuente al sistema de almacenamiento).

Los lagos de datos son "... a central location in which to store all your data, regardless of its source or format ..." (Thusso, A. & Sharma, B., 2015, p. 1), es decir, son un lugar de almacenamiento de datos centralizado independiente de la fuente y el formato, que permite almacenar los datos de una manera estructurada y no estructurada.

Thusso y Sharma (2015) p.2, plantean un modelo de lago de datos (datalake) de 5 zonas basados en el modelo de Zaloni, estas 5 zonas son:

Zona transitoria (*transient*): "(...) it is a landing zone for data where security measures can be applied before it is stored or accessed [es una zona de aterrizaje para datos donde se pueden aplicar medidas de

seguridad antes de almacenarlos o acceder a ellos]", es decir, es una zona en la cual se almacenan los datos en crudo y se emplea como zona de aterrizaje de los datos para su posterior procesamiento.

Zona cruda (raw): "(...) *In the Raw Zone, data is stored permanently and in its original form, so it is known as "the single source of truth* [en la zona cruda, los datos se almacenan de forma permanente y en su forma original, por lo que se conoce como "la única fuente de verdad"]", es decir, esta zona corresponde con los datos ya formateados (normalmente en formato parquet y orc), además de aquellos campos y registros filtrados en la zona transitoria por motivos de seguridad (por ejemplo: datos sensibles).

Zona de confianza (trusted): "(...) *The Trusted Zone is based on raw data in the Raw Zone, which is the "single source of truth."* It is altered in the Trusted Zone to fit business needs and be in accordance with set policies [La zona de confianza se basa en datos sin procesar en la zona sin procesar, que es la "fuente única de la verdad". Se modifica en la zona de confianza para adaptarse a las necesidades de la entidad y en concordancia con las políticas establecidas (gobierno de datos)]", en esta zona se almacenan todos los datos en estructuras definitivas, haciendo las veces de una bodega de datos lógica. Es decir, los datos están ya cruzados, lo que permite realizar consultas para el usuario final y la conexión a plataformas de análisis de datos o similares.

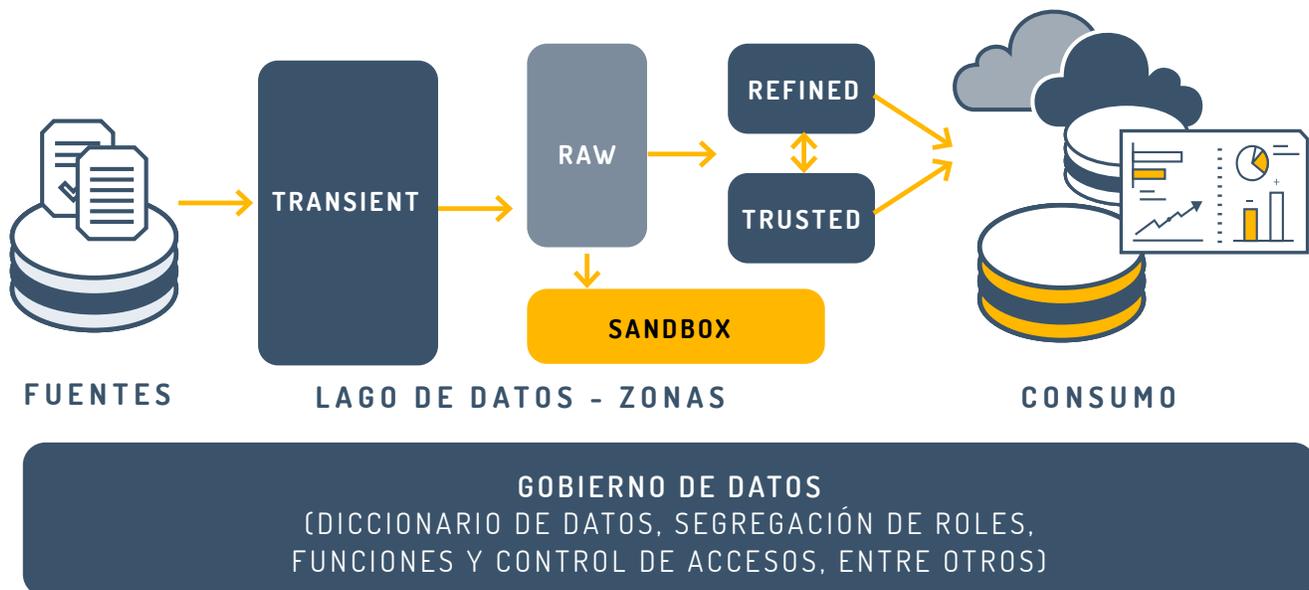
Zona refinada (refined): "(...) *Data here is integrated into a common format for ease of use, and goes through possible detokenization, further quality checks, and life cycle management* [En esta zona, los datos se integran en un formato común para facilitar su uso y pasan por una posible eliminación de identificadores, controles de calidad adicionales y gestión del ciclo de vida]", esta zona se suele emplear para encapsular información específica de cierta área o proceso y hace las veces de almacén de datos. Se suele conectar a sistemas transaccionales a través de la construcción de API's y a plataformas de análisis de datos para visualización y reporte.

Caja de arena (sandbox): "... *The Sandbox is integral to a data lake because it allows data scientists and managers to create ad hoc ex-*

ploratory use cases without the need to involve the IT department or dedicate funds to creating suitable environments within which to test the data [esta zona es parte integral del lago de datos debido a que permite a los científicos y administradores de datos crear casos de uso exploratorios inmediatos sin la necesidad de involucrar al departamento de tecnología o realizar inversiones para crear entornos adecuados dentro de los cuales probar los datos]”, es decir, esta zona suele emplearse como repositorio temporal para el cruce de información y descubrimiento de datos. Además, puede ser empleada para el entrenamiento de modelos matemáticos (principalmente de inteligencia artificial) y como repositorio personal de aquellos usuarios con acceso a consulta y cruce de datos.

En la Figura 3 se puede apreciar el flujo de datos de acuerdo con las zonas de un lago de datos. El gobierno de datos es transversal a todo el proceso e incluye la gestión de la metadata y seguridad, la adecuada segregación de roles y funciones, y una adecuada asignación de accesos, tanto para los usuarios como para las aplicaciones que requieran acceder al lago de datos.

Figura 3. Representación del flujo de datos de acuerdo con las zonas de un lago de datos.



Nota. Adaptado por Función Pública (2022) con base en Thusso, A. & Sharma, B. (2015) p. 2.

Recomendaciones en la implementación del *big data* (datos masivos)

Al momento de implementar *big data* al interior de la entidad debemos prestar atención a:

El gobierno de los datos: incluye definiciones claras de acceso, roles y responsabilidades, estructuras y procesos alrededor de la ingesta, procesamiento y uso de los datos.

Las plataforma y flujos de datos: definir claramente cuál será la plataforma de almacenamiento que se va a emplear y cuál será su infraestructura (localmente o nube pública, privada, híbrida). Por otra parte, es necesario definir la malla de procesamiento y los mecanismos que se van a emplear (incluye el lenguaje de programación) para el paso entre zonas.

Los productos generados a partir de los datos: cuáles serán los productos que se generarán a partir de los datos y cómo serán actualizados y expuestos para que los usuarios y las aplicaciones puedan acceder a ellos. Por ejemplo: modelos matemáticos, datos agregados, cálculos e indicadores, entre otros.

La gestión del cambio: es necesario que se promueva el correcto entendimiento y uso en el día a día de las herramientas y productos generados desde *big data* (datos masivos) y analítica.

Anexo C. Tipos de visualizaciones de datos

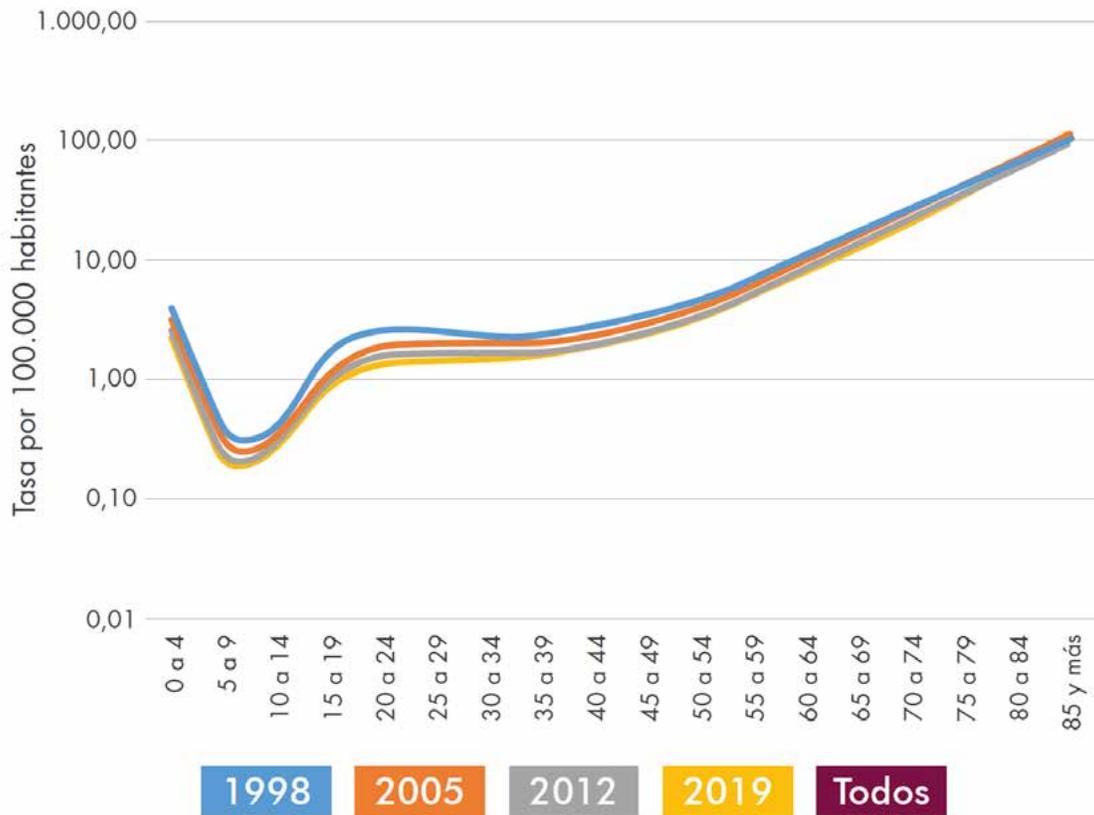
La visualización de datos es uno de los productos de la analítica institucional más empleados, debido a su facilidad para comunicar que sucede y porque sucede a través de los datos e indicadores. En este anexo, se mostrarán los principales tipos de visualizaciones de datos:

Gráfico lineal

El gráfico lineal o gráfico de serie de tiempo se emplea para analizar cómo cambia una variable en el tiempo e identificar visualmente su tendencia, estacionalidad y desviaciones a lo largo de un periodo (Wexler S. et al., 2017, pág. 40).

En la Figura 1 se aprecia un gráfico de series de tiempo que muestra la evolución de las tasas específicas de mortalidad por cada 100.000 habitantes según el grupo etario del fallecido para Colombia en los periodos 1998, 2005, 2012, 2019, y total. Donde todos los periodos comparten la misma tendencia creciente a partir de los 10 años.

Figura 1. Ejemplo de gráfico lineal por tasas específicas de mortalidad (TEM), según el grupo etario del fallecido, Colombia, 1998-2005-2012-2019



Nota. Tomado por Función Pública (2022) del Departamento Administrativo Nacional de Estadística (DANE), 2019.

Lo anterior es útil cuando se desea comparar la evolución en el tiempo de una variable por distintas características o dimensiones, permitiendo entender de una manera no rigurosa, si existe algún tipo de correlación, y si esta se mantiene en el tiempo.

Algunos ejemplos en los que se pueden utilizar este tipo de gráficos son:

- Número de ciudadanos atendidos por hora.
- Total de facturas radicadas por periodo de tiempo.
- Total de ingresos a la página web de nuestra entidad por hora.

- Total de reproducciones (video) en la web de nuestra entidad por semana.
- Total de niños, niñas, adolescentes y jóvenes atendidos por mes.
- Total de niños que han superado el estado de desnutrición (por déficit o exceso) por mes en una región específica del país.
- Total de consultas realizadas en el portal de atención al ciudadano por día.
- Porcentaje mensual de avance en los proyectos.
- Evolución de la evaluación anual de desempeño.

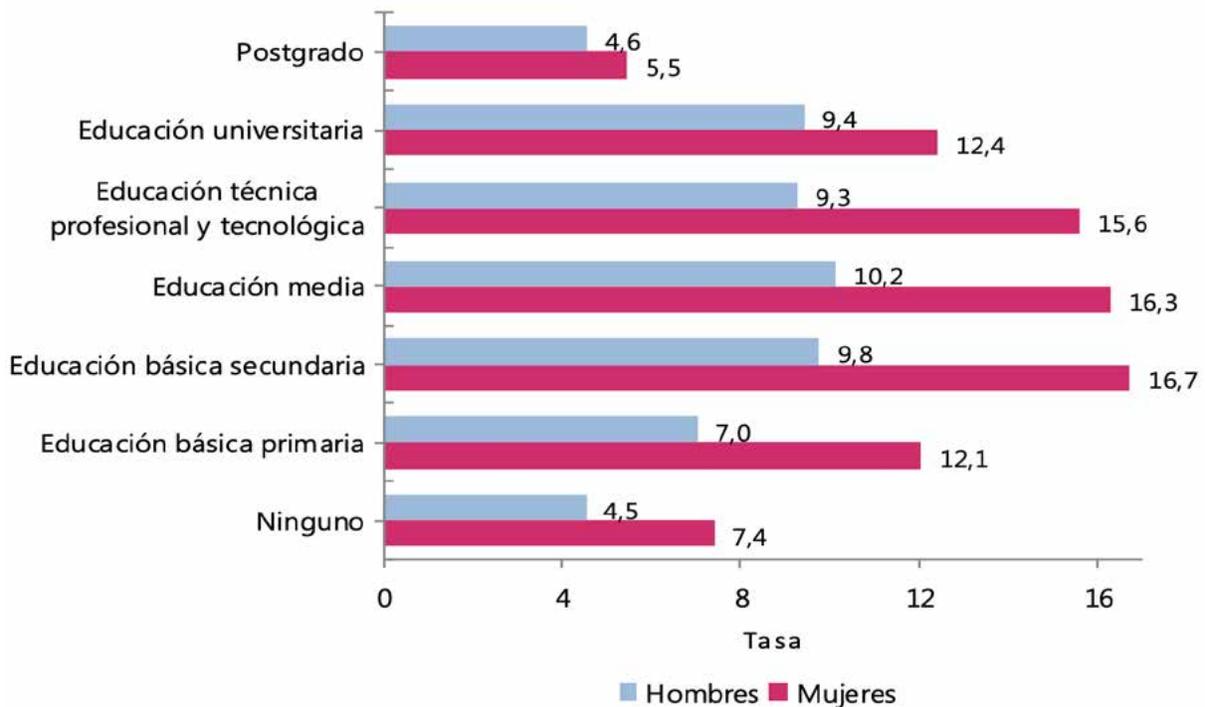
Gráfico de barras

De acuerdo con Wexler S. *et al.* (2017), se utilizan para comparar valores entre distintas categorías, lo que permite identificar pequeñas desviaciones por medio de la comparación en la longitud de las barras.

Las barras pueden estar orientadas de forma vertical u horizontal. La forma vertical es recomendada cuando existen muchas categorías (más de 10) o sus nombres son muy largos (es cortado por el *software* utilizado). En cambio, la forma horizontal se recomienda cuando la categoría corresponde a una medida de tiempo, por ejemplo: día, mes, trimestre, semestre, año.

En la Figura 2 se aprecia la tasa de desempleo por nivel educativo y sexo (total nacional) para el año 2019. En esta gráfica se identifican tasas superiores de desempleo para mujeres en todos los niveles educativos. Esto permite identificar posibles necesidades de apoyo o elaborar políticas enfocadas en esta población.

Figura 2. Ejemplo de gráfico de barras – tasa de desempleo según nivel educativo logrado y sexo (total nacional 2019)



Nota. Tomado por Función Pública (2022) del Departamento Administrativo Nacional de Estadística (DANE), 2019.

Otros ejemplos en los que se pueden utilizar este tipo de gráficos son:

- Total de atenciones al ciudadano por medio (presencial, virtual, telefónico) durante el mes actual.
- Total de niños atendidos en el programa de atención a la primera infancia por municipio para el presente año.
- Total de ciudadanos por departamento.
- Total de gastos por área o dependencia de la entidad.
- Total de inversión por municipio durante el cuatreno.

Se debe tener en cuenta que, si no es posible visualizar los valores de las barras, hay que incluir el valor numérico en ellas o líneas de división primarias y secundarias (líneas verticales y horizontales

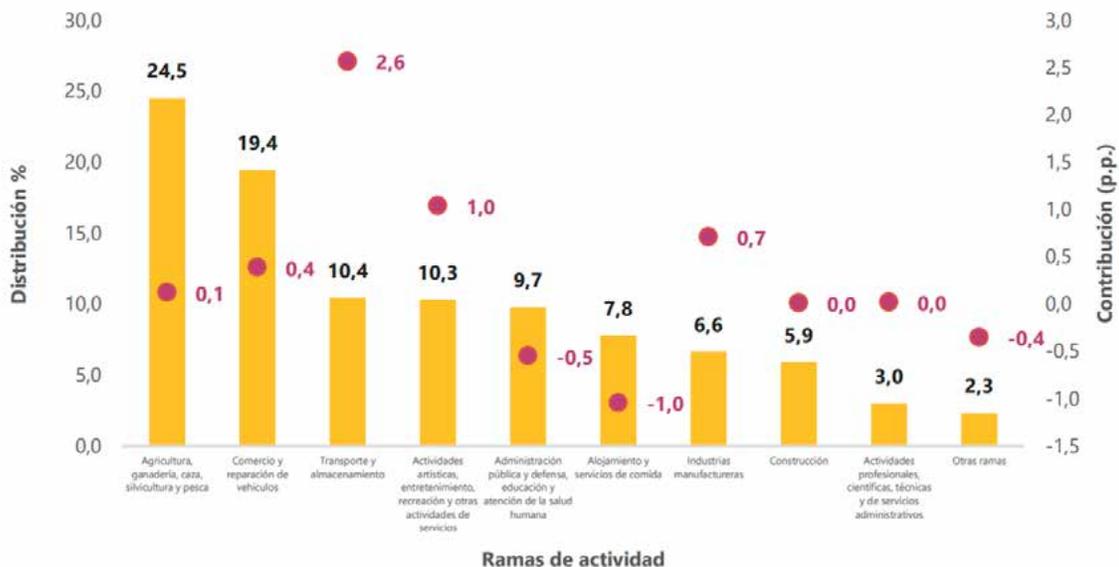
en forma de cuadrícula) que permitan tener una aproximación más cercana al valor.

Gráfico combinado

Como su nombre lo indica, corresponde a la mezcla entre distintos gráficos. Wexler S. et al. (2017) sugieren su uso cuando es necesario comparar indicadores cuyas escalas son muy distintas o se tiene un indicador base de comparación, el cual dicta un nivel de referencia para los demás.

En la Figura 3, se aprecia la distribución en barras de la tasa global de participación, ocupación y desempleo para el departamento de Sucre durante el año 2021, en la que las barras amarillas dan cuenta de la distribución en porcentaje, mientras que los puntos morados a la contribución (punto porcentual [p.p.]), permitiendo así conocer cuál es el desempeño por cada una de las ramas de actividad y saber cuánto aportan al valor global.

Figura 3. Ejemplo de gráfico combinado: tasa global de participación, ocupación y desempleo para el departamento de Sucre durante el año 2021. Las barras amarillas corresponden con la distribución (%) y los puntos morados con la contribución (p.p.)



Nota. Tomado por Función Pública (2022) del Departamento Administrativo Nacional de Estadística (DANE), 2021.

Otros ejemplos en los que se pueden utilizar este tipo de gráficos son:

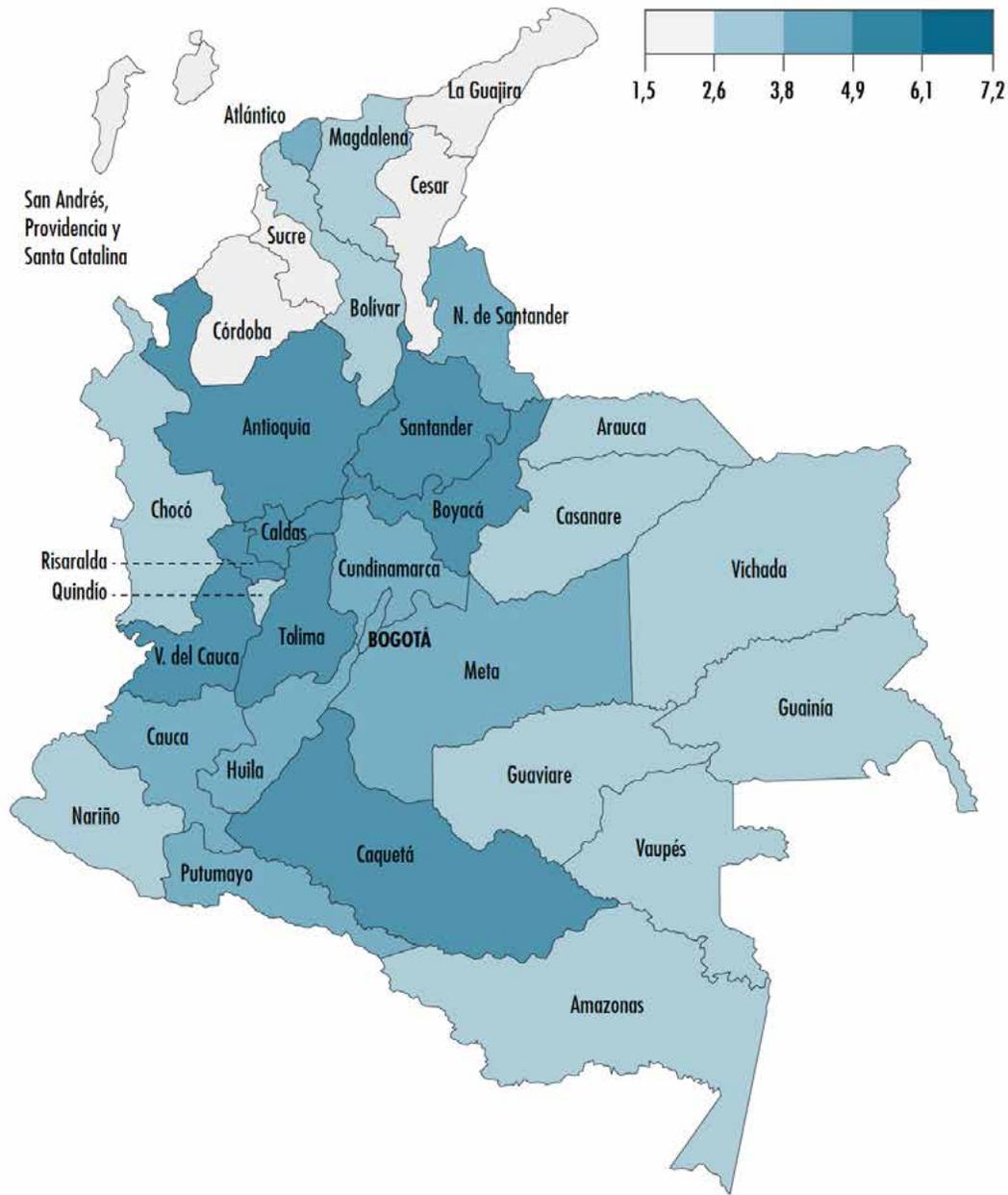
- Porcentaje mensual de satisfacción del ciudadano con la atención prestada por la dependencia comparado con la satisfacción del ciudadano total de nuestra entidad.
- Total de niños atendidos anualmente en el programa de atención a la primera infancia por municipio comparado con el total de niños del municipio.
- Total de ciudadanos por departamento comparado con el total de ciudadanos por debajo de la línea de pobreza.
- Evolución del índice de pobreza multidimensional del municipio comparado con el porcentaje de analfabetismo en el municipio.
- Total de ciudadanos impactados positivamente por el acompañamiento institucional comparado con el índice de pobreza multidimensional en el departamento.

Mapas de calor y puntos

Estos gráficos permiten mostrar el valor de un indicador sobre un área geográfica. Según Wexler S. et al. (2017), son muy útiles cuando las características propias de la región inciden en el valor del indicador o cuando se requiere identificar cobertura en el espacio.

En la Figura 4, se aprecia en colores la tasa bruta de mortalidad por cada 1.000 habitantes de acuerdo con el departamento de residencia del fallecido, lo que permite identificar las regiones con una mayor tasa de mortalidad y enfocar esfuerzos en el análisis de las causales.

Figura 4. Ejemplo de gráfico mapa de calor y puntos: tasa bruta de mortalidad (TBM) según departamento de residencia del fallecido, Colombia, 1998, 2008 y 2019



Tasa por 1.000 hab.

Nota. Tomado por Función Pública (2022) del Departamento Administrativo Nacional de Estadística (DANE), 2019.

Otros ejemplos en los que se pueden utilizar este tipo de gráficos son:

- Localización de basuras en el municipio por hora y día.
- Ubicación de las viviendas en riesgo de un movimiento lineal en masa en el municipio.
- Cantidad de hechos delictivos por día y zona en el municipio.
- Concentración de ciudadanos vulnerables por barrio.
- Capacidad utilizada de urgencias en los centros médicos del municipio.
- Tasas de mortalidad por departamento.
- Tiempos de desplazamiento entre zonas del municipio.
- Nivel de material particulado en el ambiente por zona del municipio.
- Nivel fluvial por trayecto de cada fuente hídrica del municipio y región.
- Cantidad de hechos delictivos por zona del municipio.
- Concentración de instituciones educativas por zona del municipio.

Gráfico de dispersión y burbujas

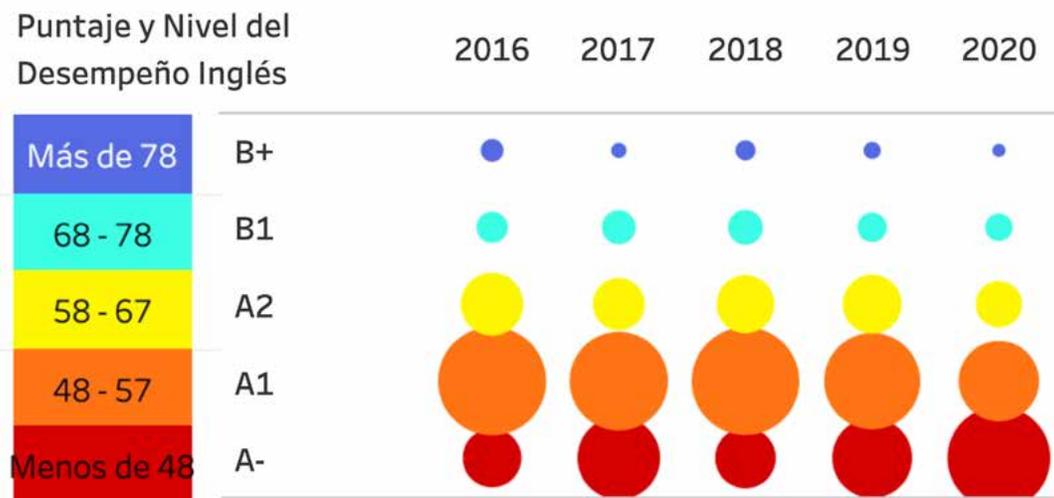
Para Wexler S. et al. (2017), este tipo de gráfico es muy útil cuando es necesario combinar varias dimensiones, categorías o atributos en un único gráfico, para lo cual se pueden asociar al tamaño, el color y la posición de los mismos, siendo este último la intersección de dos valores numéricos en los ejes vertical y horizontal.

La forma de construirlo comienza con la intersección de dos valores numéricos, uno para el eje horizontal (X) y otro para el eje vertical (Y); a partir de allí se puede agregar una categoría a los datos (correspondiente al color), por ejemplo: departamento, ciudad, proceso, área o dependencia, tipo de atención, entre otros.

Adicionalmente, es posible agregar un valor numérico que dé cuenta del tamaño de la burbuja. En algunos casos también es posible agregar una línea de tiempo para ver su cambio y evolución, para lo cual es necesario asociarle una variable ordenada que dé cuenta de la fecha o alguna medida de tiempo.

En la Figura 5 se muestra un ejemplo de este gráfico en el que se puede identificar la relación que existe entre el nivel de inglés por año (color) y el porcentaje de establecimientos educativos en ese puntaje (tamaño del círculo) por año, notándose una disminución en puntaje A1 y aumentando en A- para el 2020.

Figura 5. Ejemplo de gráfico de dispersión y burbujas: ¿Cuánto sabemos de inglés en Medellín?, acercamiento al estado del aprendizaje de inglés en la ciudad



Nota. Tomado por Función Pública (2022) del portal de datos abiertos de Medellín, MEDATA, 2021.

Otros ejemplos pueden ser:

- Índice de pobreza multidimensional por edad (ejes) por sexo (color) y cantidad de población (tamaño de la burbuja).
- Relación del nivel de lluvias con la cantidad de deslizamientos (ejes) por zona (color) y total de hogares afectados (tamaño de la burbuja).

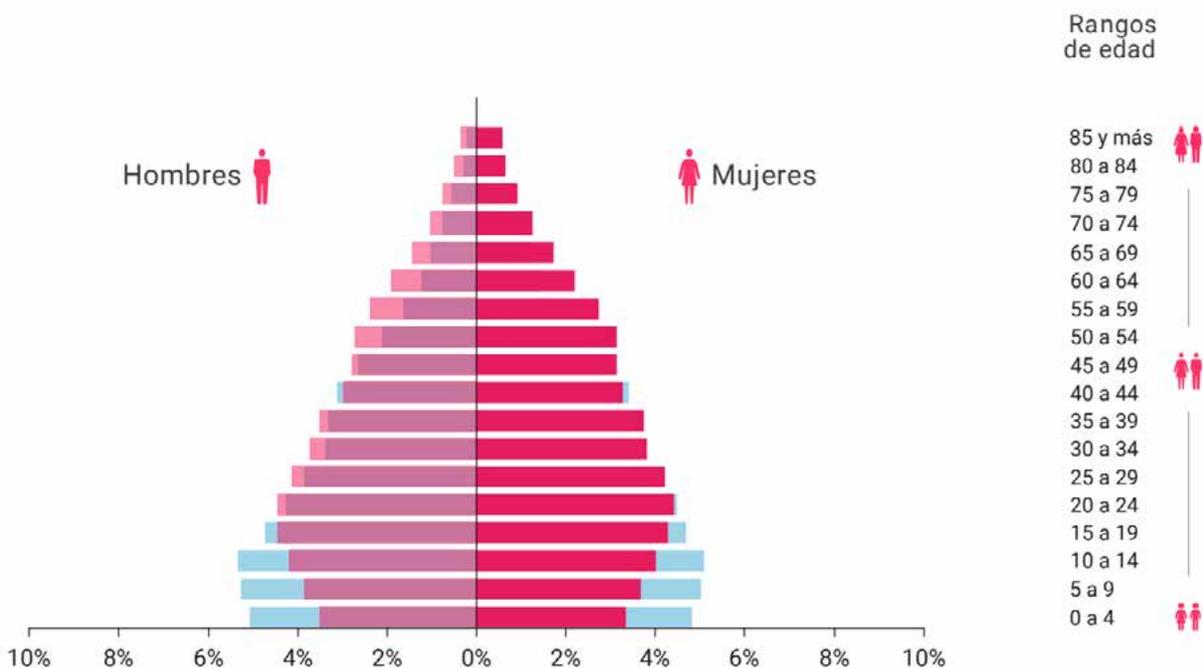
- Índice de pobreza multidimensional por estado de nutrición de los niños del municipio (ejes) por edad (color) y cantidad de niños (tamaño de la burbuja).

Gráfico de cascada y embudo

Para Wexler S. et al. (2017), corresponden a aquellos gráficos que permiten identificar por qué se logra un valor del indicador a partir de sus componentes (cascada) y entender cómo cambia por distintos rangos, principalmente de tiempo (embudo).

En la Figura 6 se muestra la distribución poblacional por rango de edad en 2018 para Colombia por sexo, entendiendo el aporte al total por cada rango de edad y sexo.

Figura 6. Ejemplo de gráfico de cascada y embudo – distribución población colombiana a partir del censo 2018



Nota. Tomado por Función Pública (2022) del Departamento Administrativo Nacional de Estadística (DANE), 2018.

Otros ejemplos en los que se pueden utilizar este tipo de gráficos son:

- Retorno de la inversión de los proyectos a partir de las principales líneas de ingreso y costo.
- Total de ciudadanos por rango de edad por año, mostrando el aporte de los nacimientos y defunciones.
- Tiempo total de atención al ciudadano a partir de la medición del tiempo en cada paso de atención.
- Satisfacción de los ciudadanos en la atención de un PQRS.
- Cantidad de niños menores de cinco años en estado de desnutrición por causa.

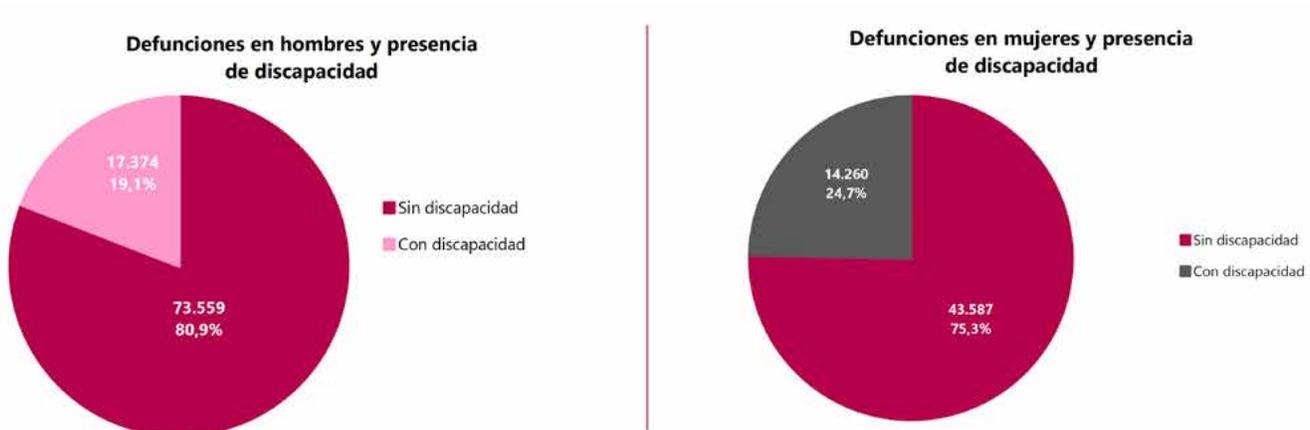
Gráficos circulares y de anillos

Para Wexler S. *et al.* (2017), los gráficos circulares concéntricos pueden tener un centro blanco (anillos) o lleno (circulares o de torta), en los que se requiere analizar la proporción de una categoría sobre el total. Se debe tener en cuenta que cuando existen muchas categorías es mejor usar un gráfico de barras o similar.

Es posible generar anillos exteriores adicionales, en cuyo caso se recomienda no exceder una cantidad visualmente agradable de ellos (3).

En la Figura 7 se muestra un ejemplo en el que se identifica la distribución de las defunciones por COVID-19 según sexo y presencia de discapacidad para Colombia del 16 de marzo de 2020 al 6 de marzo de 2022 preliminar, donde es posible entender los totales de cada uno y el porcentaje sobre el total con el fin de hacer comparaciones y análisis relativos.

Figura 7. Ejemplo de gráfico circular y de anillos: distribución de las defunciones por COVID-19 según sexo y presencia de discapacidad total nacional: 16 de marzo 2020 al 6 de marzo 2022 preliminar



Nota. Tomado por Función Pública (2022) del Departamento Administrativo Nacional de Estadística (DANE), 2020.

Algunos ejemplos en los que se pueden utilizar este tipo de gráficos son:

- Distribución de población por estrato.
- Distribución de población por sexo.
- Distribución de casos por tipología.
- Distribución de gastos por área o dependencia.

Tablas

Las tablas se utilizan para representar distribuciones de valores en filas y columnas. Según Wexler S. *et al.* (2017), es necesario tener en cuenta qué información es relevante e intercalarla con otros tipos de elementos visuales.

En la Figura 8 se muestra un ejemplo de las tasas específicas de mortalidad por sexo y la razón de sexos de la mortalidad, permitiendo identificar qué está sucediendo por sexo en un registro por años.

Figura 8. Ejemplo de tablas tasas específicas de mortalidad (TEM) por sexo y razón de sexos de la mortalidad: Colombia, 1998-2019

Año	TEM Hombres	TEM Mujeres	Razón de sexos de la mortalidad
1998	5,7	3,6	1,6
1999	5,8	3,7	1,6
2000	5,9	3,7	1,6
2001	5,9	3,8	1,6
2002	5,9	3,7	1,6
2003	5,7	3,8	1,5
2004	5,5	3,7	1,5
2005	5,4	3,7	1,4
2006	5,4	3,8	1,4
2007	5,3	3,8	1,4
2008	5,3	3,8	1,4
2009	5,3	3,8	1,4
2010	5,3	3,8	1,4
2011	5,1	3,7	1,4
2012	5,1	3,8	1,4
2013	5,2	3,8	1,4
2014	5,2	4,0	1,3
2015	5,4	4,2	1,3
2016	5,4	4,2	1,3
2017	5,4	4,2	1,3
2018	5,6	4,3	1,3
2019	5,6	4,3	1,3

***Las Tasas Específicas de Mortalidad - TEM se presentan por 100.000**

Nota. Tomado por Función Pública (2022) del Departamento Administrativo Nacional de Estadística (DANE), 2019.

En general, pueden emplearse con cualquier tipo de dato, pero su fortaleza radica cuando queremos entender el detalle o las cifras cambian en muy poco tiempo, por ejemplo:

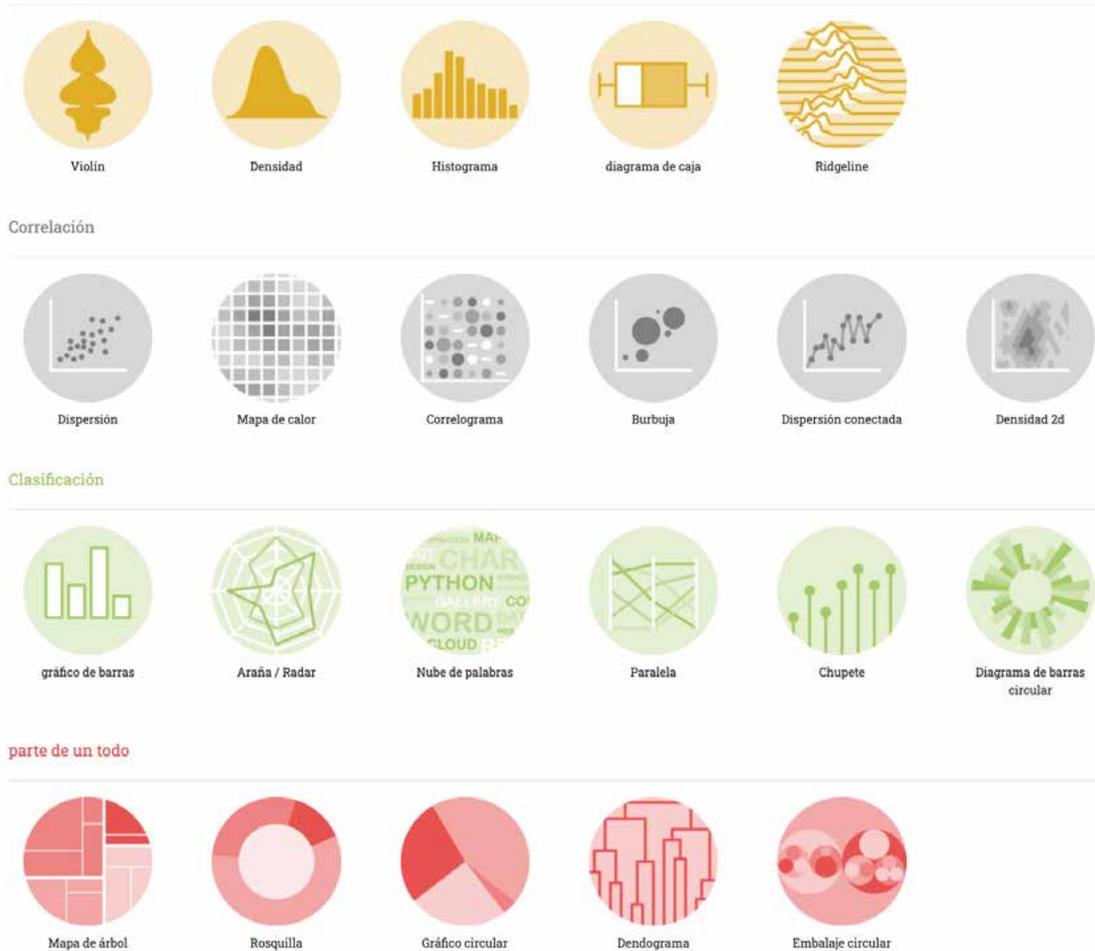
- Tasas de cambio específicas de mortalidad agrupadas por sexo y edad.
- Satisfacción por grupos de ciudadanos en un proceso específico.

- Valores antropométricos de una población y su evolución año a año.
- Variación porcentual en la medición de un indicador.
- Tasas promedio de crecimiento económico en una región.

Otros tipos de gráficos

Podemos crear otros tipos de gráficos en los softwares empleados o hacer uso de herramientas abiertas digitales como D3.js, que ofrece un conjunto de gráficos desarrollados para uso libre en múltiples lenguajes. Algunos de los tipos se pueden apreciar en la Figura 9 y en su página web: <https://www.d3-graph-gallery.com>

Figura 9. Ejemplos de tipos de gráficos disponibles en d3-graph



Nota. Tomado de D3.js Graph Gallery.



Bibliografía

Alcaldía de Medellín (2022). Análisis de embarazos en adolescentes. <http://medata.gov.co/medell%C3%ADn-en-cifras/an%C3%A1lisis-de-embarazo-en-adolescentes>

Alcaldía de Medellín (2021). Inglés en Medellín. <https://www.medellin.gov.co/pmu/#/tablero/2>

Brockwell, P. & Davis, R. (2016). Introduction to Time Series and Forecasting, Springer International Publishing, New York.

Davenport, T.H. & Prusak, L. (1998). Working Knowledge. Harvard Business School Press, Boston.

Departamento Administrativo Nacional de Estadística (2018). Infografía del censo nacional de población y vivienda 2018-Colombia. <https://www.dane.gov.co/files/censo2018/infografias/info-CNPC-2018total-nal-colombia.pdf>

Departamento Administrativo Nacional de Estadística (2019). Anuario nacional de estadísticas vitales. <https://www.dane.gov.co/files/investigaciones/poblacion/anuario-EEVV-2019/anuario-nacional-de-estadisticas-vitales-colombia-2019.pdf>

Departamento Administrativo Nacional de Estadística (2022). Estadísticas vitales- EEVV. https://www.dane.gov.co/files/investigaciones/poblacion/pre_estadisticasvitales_IVtrim_2021pr.pdf

Departamento Administrativo de la Función Pública (2021). Guía para la analítica de datos y su uso en la planificación y ejecución de auditorías internas basadas en riesgos. <https://www.funcionpublica.gov.co/web/eva/detalle-publicacion?entryId=40840033>

Departamento Nacional de Planeación (2021). Aprovechamiento de datos para la toma de decisiones en el sector público. <https://>

www.dnp.gov.co/estudios-y-publicaciones/publicaciones/Paginas/2021.aspx

Jones, H. (2019). Minería de Datos: Guía de Minería de Datos para Principiantes, que Incluye Aplicaciones para Negocios, Técnicas de Minería de Datos, Conceptos y Más.

Kerzner, H. (2017). Project Management. Metrics, KPIs, and dashboards. A guide to measuring and monitoring project performance. Tercera edición. Nueva Jersey: John Wiley & Sons, Inc.

Kimball, R. & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modelling. Wiley.

Microsoft (2021). Vistas, <https://docs.microsoft.com/es-es/sql/relational-databases/views/views?view=sql-server-ver16>

Naeem, T. (2020, noviembre). Comprender los datos estructurados, semiestructurados y no estructurados. ASTERA. <https://www.astera.com/es/type/blog/structured-semi-structured-and-unstructured-data/>

Ortega, S. (s.f). Del Big Data al Data Quality. La gestión de la calidad de los datos. [https://blog.sortega.com/del-big-data-al-data-quality-la-gestion-de-la-calidad-de-los-datos/#:~:text=Exactitud%20\(Accuracy\)%3A%20Se%20mide,un%20evento%20que%20se%20describe](https://blog.sortega.com/del-big-data-al-data-quality-la-gestion-de-la-calidad-de-los-datos/#:~:text=Exactitud%20(Accuracy)%3A%20Se%20mide,un%20evento%20que%20se%20describe)

Perez, C. (2022). Data Mining. The CRISP-DM Methodology. The CLEM language and IBM SPSS Modeler.

Ponniah, P. (2011). Data Warehousing Fundamentals for IT Professionals. Wiley.

Rodríguez, P. et al. (2017). El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe, Banco Interamericano de Desarrollo.

Salazar, C & Del Castillo, P. (2018). Fundamentos básicos de estadística. 2018. <http://www.dspace.uce.edu.ec/bitstream/25000/13720/3/Fundamentos%20B%20C3%A1sicos%20de%20Estad%20C3%ADstica-Libro.pdf>

Shearer C. (2000). The CRISP-DM model: the new blueprint for data mining. J Data Warehousing.

Sherman, R. (2015). Business Intelligence Guidebook: From Data Integration to Analytics. Morgan Kaufmann.

Sinergia (s.f.). Bases de datos OLTP y OLAP, https://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx#:~:text=Los%20sistemas%20OLAP%20son%20bases%20de%20datos%20orientadas%20al%20procesamiento%20anal%20C3%ADtico

Technics Publications. [Ed.]. (2017), DAMA-DMBOK, guía del conocimiento para la gestión de datos, segunda edición. Estados Unidos: Editorial Dama International.

Thusso, A. & Sharma, B. (2015). Architecting Data Lakes, O'Reilly Media, Inc.

Tshilidzi, M. (2015). Causality, Correlation and Artificial Intelligence for Rational Decision Making. World Scientific.

Wexler, S. et al., (2017). The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios. Wiley.



FUNCIÓN PÚBLICA

Analítica institucional: el poder de los datos en el sector público

Versión 1

SEPTIEMBRE DE 2022

Dirección de Gestión del Conocimiento

Función Pública

Carrera 6 n.º 12-62, Bogotá, D.C., Colombia

Conmutador: (+57) 601 7395656

Fax: (+57) 601 7395657

Web: www.funcionpublica.gov.co

eva@funcionpublica.gov.co

Bogotá, D.C., Colombia.

VISÍTANOS O ESCRÍBENOS:

