Módulo 3. Calidad de la data

Unidad 3.1

Objetivos del módulo

1. Entender la importancia y trascendencia que tiene la calidad de la data en una organización.
2. Identificar los diferentes actores que intervienen en la consecución de la calidad de la data en una organización.
3. Entender cómo la calidad de la data lleva a las organizaciones a una cultura *data driven*.

Introducción

Muchas organizaciones piensan que simplemente porque generan muchos informes o tienen muchos paneles, están basados en datos. Aunque esas actividades son parte de lo que hace una organización, también son típicamente retrógradas. Es decir, a menudo son una declaración de hechos pasados o presentes sin mucho contexto, sin explicación causal de por qué algo ha sucedido y sin recomendaciones de qué hacer a continuación. En resumen, dicen lo que pasó, pero no son prescriptivos.

Lograr tales conocimientos requiere la recopilación de los datos adecuados y dignos de confianza; que el análisis sea bueno; que los conocimientos se consideran en la decisión e impulsen acciones concretas para que el potencial se pueda alcanzar.

La calidad de la data tiene un significado de gran alcance y es relevante, ya que la ausencia de ella tiene grandes consecuencias en la eficiencia y efectividad en las organizaciones y los negocios. Las consecuencias de la mala calidad de los datos se experimentan a menudo en la vida cotidiana, pero sin hacer las conexiones necesarias con sus causas. El objetivo de este módulo está orientado a introducir las perspectivas relevantes que hacen de la calidad de los datos un tema que vale la pena ser investigado y comprendido.

- Comparación de clientes: los sistemas de información de las organizaciones públicas y privadas pueden considerarse el resultado de un conjunto de actividades escasamente controladas e independientes que producen varias

bases de datos, muy a menudo caracterizadas por la superposición de la información. En organizaciones privadas, no es sorprendente tener varios registros de clientes, actualizados con diferentes procedimientos organizativos, lo que resulta en información inconsistente y duplicada.

- Fusión de organizaciones: cuando diferentes organizaciones o diferentes unidades de una organización se fusionan, es necesario integrar los sistemas de información heredados. Tal integración requiere compatibilidad e interoperabilidad en cualquier capa del sistema de información, con el nivel de base de datos requerido para garantizar la interoperabilidad física y semántica.

Los ejemplos anteriores son indicativos de la creciente necesidad de integrar la información a través de fuentes de datos completamente diferentes, una actividad en la que la pobre calidad de los datos dificulta los esfuerzos de integración.

La calidad de los datos es más que un asunto tecnológico. Requiere que la alta dirección trate los datos como un activo y darse cuenta de que el valor de este activo depende de su calidad.
La calidad de los datos es un concepto multifacético, en cuya definición coinciden diferentes dimensiones. Las dimensiones de calidad —por ejemplo, la precisión— pueden detectarse fácilmente en algunos casos (por ejemplo, errores ortográficos), pero son más difíciles de detectar en otros casos (por ejemplo, cuando se proporcionan valores admisibles, pero no correctos).

3.1.1 Calidad de la data y tipos de data

Los datos representan objetos del mundo real en un formato que puede ser almacenado, recuperado y elaborado por un *software*; y comunicado a través de una red. El proceso de representar el mundo real por medio de datos puede aplicarse a un gran número de fenómenos, tales como mediciones, eventos, características de las personas, el medio ambiente, sonidos y olores. Los datos son extremadamente versátiles en dicha representación

Además de los datos, otros tipos de información son utilizados en la vida real y los procesos de negocio, como la información en papel y la información transmitida por la voz. No vamos a tratar con todos estos tipos de información, nos concentraremos en los datos.

Dado que los investigadores en el campo de la calidad de los datos deben tratar con un amplio espectro de posibles representaciones de datos, han propuesto varias clasificaciones para los datos. En primer lugar, varios autores distinguen, implícita o explícitamente, tres tipos de datos:

1. Estructurado: cuando cada elemento de datos tiene una estructura fija asociada. Las tablas relacionales son el tipo más popular de datos estructurados.

2. Semiestructurado: cuando los datos tienen una estructura que tiene cierto grado de flexibilidad. XML es el lenguaje de marcado comúnmente utilizado para representar datos semiestructurados. Algunas características comunes de los datos semiestructurados son:

> i) Los datos pueden contener campos no conocidos en el momento del diseño, por ejemplo, un archivo XML no tiene un archivo de esquema XML asociado.
> ii) El mismo tipo de datos puede representarse de múltiples maneras, por ejemplo, una fecha puede ser representada por un campo o por varios campos, incluso dentro de un único conjunto de datos.
> iii) Entre los campos conocidos en el momento del diseño, muchos campos no tendrán valores.

3. No estructurados: cuando los datos se expresan en lenguaje natural y no se definen estructuras o tipos de dominio específicos.

En este módulo, hemos percibido que la calidad de los datos es un área multidisciplinaria. Esto no es sorprendente, ya que los datos, en una variedad de formatos y con una variedad de los medios de comunicación, se utilizan en toda actividad de la vida real o de negocios; e influyen profundamente en la calidad de los procesos que utilizan los datos. Muchas organizaciones públicas y privadas han percibido el impacto de la calidad de los datos en sus activos y misiones, por consiguiente, han puesto en marcha iniciativas de gran impacto. Al mismo tiempo, mientras que, en los sistemas de información monolíticos, los datos se procesan dentro de actividades controladas, con la llegada de las redes e Internet, los datos se crean e intercambian con procesos mucho más "turbulentos" y necesitan una gestión más sofisticada.

Si bien la calidad de los datos es una esfera de investigación relativamente nueva, otras esferas —como el análisis de los datos estadísticos— han abordado en el pasado algunos aspectos de los problemas relacionados con la calidad de los datos. El análisis de datos estadísticos, la representación de conocimientos, la extracción de datos, los sistemas de información de gestión y la integración de datos comparten algunos de los problemas y cuestiones característicos de la calidad de los datos y, al mismo tiempo, proporcionan paradigmas y técnicas que puedan utilizarse eficazmente en las actividades de medición y mejora de la calidad de los datos.

"80 % of my time was spent cleaning the data. Better data will always beat better models" (Thomson Nguyen, 2015).

Traducción: "El 80 % de mi tiempo lo dediqué a limpiar los datos. Los mejores datos siempre les ganan a los mejores modelos" (Thomson Nguyen, 2015)

Los datos son la base de una organización basada en datos. Si no tiene datos oportunos, relevantes y confiables, los tomadores de decisiones no tienen otra alternativa que tomar decisiones por intuición. La calidad de los datos es clave.
Los analistas necesitan los datos correctos, recopilados de la manera correcta, en la forma correcta, en el lugar correcto y en el momento adecuado —y no están pidiendo mucho—. Si alguno de estos aspectos falta, los análisis son limitados en cuanto a las preguntas que pueden responder; y el tipo o la calidad de las percepciones que pueden derivar de los datos.

3.1.2 Facetas en la calidad de la data

La calidad de los datos no es algo que se pueda reducir a un solo número. La calidad no es un 5 o un 32. La razón es que el término cubre un conjunto de facetas o dimensiones. En consecuencia, hay grados de calidad, con algunos problemas más graves que otros. Sin embargo, la gravedad de esas cuestiones puede depender del contexto del análisis que se realice con los datos. Los datos tienen un número de facetas. La data deberá ser:

- ➔ Accesible

Un analista tiene acceso a los datos. Esto no solo cubre los permisos, sino también las herramientas apropiadas que hacen que los datos sean utilizables y analizables

La data tiene que estar expuesta en una base de datos en funcionamiento o en una herramienta de inteligencia empresarial para que los analistas puedan analizar los datos.

- ➔ Precisa

Los valores representan el verdadero valor o estado de la entidad. Por ejemplo, un termómetro mal calibrado, una fecha de nacimiento mal escrita o una dirección de cliente desactualizada representan datos inexactos.

- ➔ Coherente

Los datos se pueden combinar con otros datos relevantes en una manera precisa. Por ejemplo, un pedido de venta debe poder estar vinculado a un cliente; uno o más productos en el pedido; una dirección de facturación y/ o envío; y, posiblemente, información de pago. Ese conjunto de información proporciona una

imagen coherente del orden de venta. La coherencia es impulsada por el conjunto de claves que unen los datos en diferentes partes de la base de datos.

➔ Completa

No hay datos que falten. Esto puede significar una sola pieza de datos dentro de un solo registro, como un nombre que falta en un registro de cliente; o registros completos que faltan, como un registro de cliente completo que no se pudo guardar en una base de datos.

➔ Consistente

Los datos son un agregado. Por ejemplo, una dirección de correo electrónico para un cliente en particular en una fuente de datos coincide con la dirección de correo electrónico del mismo cliente en otra fuente de datos. Cuando hay conflictos, una fuente debe ser considerada la fuente maestra o ambas no son utilizadas hasta que la fuente de desacuerdo sea entendida y corregida.

➔ Definida

Los campos de datos individuales tienen un significado bien definido e inequívoco. Los campos bien nombrados y acompañados de un diccionario de datos ayudan en la calidad de la data.

➔ Relevante

Los datos guardan relación con el análisis de data que se está realizando. La información debe tener un fin determinado y deberá aportar al trabajo o proyecto que se desea llevar a cabo.

➔ Confiable

Los datos deben estar completos y se deben tener todos los datos que usted debe esperar. Si la data es exacta, los datos proporcionan la información correcta.

➔ Oportuna

Hay una duración corta o razonable entre la recopilación de los datos y la disponibilidad y liberación a los analistas. En la práctica, esto significa que los datos llegan a tiempo para que los analistas completen el análisis antes de lo previsto.

Un desvío en solo una de estas facetas puede hacer que los datos sean inútiles, parcialmente utilizables o, lo peor de todo, aparentemente utilizables pero engañosos.

3.1.3 Data sucia

Los datos pueden estar mal de muchas maneras y en cada paso del proceso de recopilación de datos. Los datos son siempre más sucios de lo que imaginamos. Según un estudio, la calidad de los datos es mala o sucia y cuesta a las empresas estadounidenses 600 000 millones de dólares anuales (¡Eso es el 3,5 % del PIB!).

En muchas situaciones, los analistas tienen poco control en la recopilación y procesamiento primario de datos. Por lo general, reciben un conjunto de datos aguas abajo en una larga cadena de pasos que abarcan la generación, el registro, la transferencia, el procesamiento y la mezcla de datos. Sin embargo, es importante conocer y apreciar los tipos de problemas de calidad de los datos que pueden surgir y sus posibles soluciones.

Por lo tanto, a partir de la fuente, ¿qué hace que los datos sean sucios y qué se puede hacer al respecto?

1. Generación de data: la generación de datos es la fuente más importante de problemas y puede surgir de errores en *hardware*, *software* y/o humanos. En el *hardware*, los sensores pueden estar mal calibrados o no calibrados, lo que puede resultar en lecturas inexactas. Por ejemplo, un sensor de temperatura podría estar leyendo alto si dice 95°F cuando en realidad es solo 93°F. Eso puede ser fácil de arreglar, porque, cuando sea posible, se debe compararlo contra alguna fuente de verdad como otro sensor o medidor de confianza, durante la configuración.

2. Entrada de la data: cuando los datos se generan manualmente, como las enfermeras que toman el peso de los pacientes, hay que registrarlos, en última instancia, en algún tipo de computadora. A pesar de la promesa de oficinas sin papel, los datos todavía se registran con demasiada frecuencia en formularios en papel como un paso intermedio antes de que se ingrese en una computadora. Estas etapas en papel pueden producir muchos errores.

En términos más generales, los problemas de entrada de datos se manifiestan en cuatro tipos de cuestiones:

- Transcripción

Las palabras o valores introducidos no son los que estaban en los datos originales.

- Inserción

Se introdujeron caracteres adicionales: 56.789 => 564.789.

- Omisión

Se omitieron uno o más caracteres: 56.789 => 56.89.

- Transposición

Se intercambiaron dos o más caracteres: 56.789 => 56.798.

Dicho esto, ¿qué podemos hacer entonces para mitigar esta clase de errores?

Lo primero que hay que hacer, si es posible, es reducir el número de pasos desde la generación de datos hasta la entrada. Para indicar lo obvio, si puede evitar un formulario de papel intermedio, debe hacerlo e ingresar los datos directamente en la computadora. Cuando sea posible, agregue validación de campo a sus formularios electrónicos. Esto asegura que los datos están bien estructurados y en el formato esperado, si no lo hace, rechace los datos y guíe al usuario para corregir cualquier error.

Cuando hay un conjunto de valores que son válidos, como abreviaturas de estado, puede usar un menú desplegable para que el usuario elija. Autocompletar es otra alternativa. En general, si desea que los usuarios escriban la menor cantidad de entradas posibles, haga que elijan entre un conjunto de opciones que usted proporcione, a menos que, por supuesto, sea una pregunta abierta con un campo de texto de forma libre. Todo esto se puede configurar y logrará hacer que la data sea lo más limpia posible.

Cuando los analistas reciben los datos, generalmente deben hacer un análisis exploratorio para evaluar la calidad de los datos.

Una manera simple de comprobar si hay errores evidentes es resumir los datos. Para cada variable, se puede calcular un resumen de cinco números: mínimo; el cuartil inferior (percentil 25); la media y/ o mediana; el cuartil superior (percentil 75); y el máximo. ¿Tienen sentido, entonces, los datos que el analista recibió? Siempre se deben hacer validaciones de lo que se recibe.

3.1.4 Procedencia de la data

Cuando se encuentran problemas de calidad de los datos, es crucial rastrearlos hasta su origen. De esa manera, todo el subconjunto de datos puede ser eliminado del análisis o mejores procesos o protocolos pueden ser diseñados y puestos en marcha para remediar el problema. Los metadatos que almacenan el origen y el historial de cambios de los datos se conocen como linaje, genealogía o —como lo estamos expresando acá— procedencia de la data.

Hay dos clases primarias de procedencia: la procedencia de la fuente, que rastrea de dónde provienen los datos; y la procedencia de la transformación, que rastrea

los cambios realizados en los datos. Por lo general, las tablas de recepción de datos sin procesar, que se llaman tablas de aterrizaje o de estadificación, tienen dos campos adicionales: tiempo de carga (el momento en que se inicia la carga de ese archivo o lote) y el nombre del archivo. De esa manera, si se descubren problemas de calidad, es muy fácil identificar de qué archivo provienen los datos para que podamos inspeccionar la línea exacta en el archivo de datos sin procesar y le pidamos al proveedor que investigue. Este es un ejemplo de procedencia.

3.1.5 La calidad de los datos es una responsabilidad compartida

Las formas en que los datos pueden ser inexactos o de mala calidad son infinitas. Además de los mencionados anteriormente, hay problemas de terminación de líneas; problemas de codificación donde los valores Unicode se comprimen (esto sucede todo el tiempo); datos corruptos; archivos truncados; datos tardíos; y nombres y direcciones que no coinciden. La calidad de los datos no debería dejarse únicamente en manos de los ingenieros de datos, debería ser una responsabilidad de toda la empresa.

Cuando los formularios de datos de entrada son parte de las operaciones de su organización, los propietarios de negocios (es decir, gerentes en las unidades de negocio), expertos en dominios y analistas deben trabajar con los desarrolladores *frontend* para proporcionar límites de verificación de rango. También deben formar parte de los requisitos y del proceso de gestión de proyectos para asegurarse de que los elementos de calidad de los datos se incorporen al proceso de flujo de datos cuando proceda. Como se mencionó anteriormente, la organización de análisis debe ser una parte interesada en el mecanismo de recopilación de datos.

3.1.6 Hacia una cultura *data driven*

Como hemos visto, la calidad de los datos tiene el potencial de garantizar la viabilidad de proyectos de *people analytics* consistentes y efectivos. Porque nos aseguramos de que lo que se mide es acorde con la realidad del fenómeno y así tener certeza de que nuestras decisiones pueden tener un impacto real en el negocio.

Pero ¿Qué sucede cuando una organización no comprende la importancia de gestionar datos de calidad o no sabe cómo hacerlo? Se puede realizar una inversión en nuevas tecnologías que administren datos en la nube, capacitar al personal en el uso de herramientas de visualización de información, entre otras iniciativas. Sin embargo, si la organización no posee una cultura de datos favorable, la calidad de los datos será el último problema en que nos deberíamos preocupar.

3.1.7 ¿A qué nos referimos con cultura de datos?

Según el portal Tableau (2021),

> La cultura de datos son las creencias y los comportamientos colectivos de las personas que valoran, aprovechan y promueven el uso de datos para mejorar la toma de decisiones. Como resultado, los datos se integran en las operaciones, la mentalidad y la identidad de una organización. Una cultura de datos permite a todos acceder a la información que necesitan para realmente basarse en los datos y superar los desafíos empresariales más complejos.

Una organización que permea una cultura de datos no solo se preocupa por conseguir datos de calidad, sino también por tener acceso a la información para la toma de decisiones en favor del negocio.

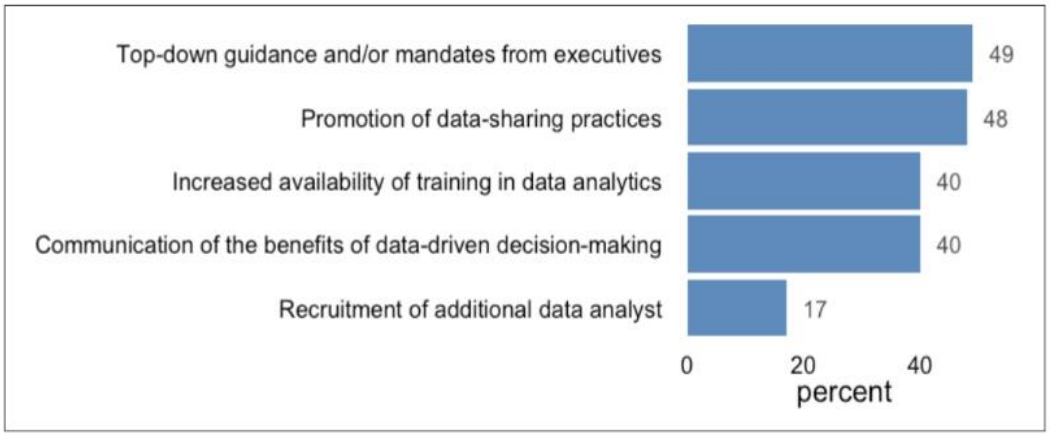
Por otro lado, es fácil imaginar que, por lo general, las distintas fuentes de datos que gestionamos son administradas por múltiples equipos o unidades comerciales diferentes. Para que la organización maximice el potencial de los datos, los datos deben combinarse para proporcionar ese contexto más completo y rico. Aquí es donde entra la cultura de datos. Tiene que haber una señal clara del negocio de que los datos no son "propiedad" de esos equipos individuales, sino que pertenecen a la organización como un todo. Los líderes de datos (discutidos más adelante) necesitan evangelizar los beneficios de compartir a la organización en su conjunto. Sin embargo, en su defecto, tienen que existir los incentivos correctos para romper los silos y compartir los datos.

Por supuesto, debe hacer todo esto sin comprometer el cumplimiento ni aumentar el riesgo. Esas son preocupaciones válidas. Un tercio de los encuestados en una encuesta de 530 ejecutivos realizada por The Economist Intelligence Unit dijo que "su empresa lucha por lograr una cultura basada en datos en parte debido a las preocupaciones sobre los problemas de privacidad y seguridad que surgen cuando se comparten datos" (como se cita en Anderson, 2015). Debido en parte a estas preocupaciones prácticas, pero, también, en parte a la inercia de que el valor predeterminado de los propietarios de negocios sea el acaparamiento de datos. Por lo tanto, esto es algo que los líderes de datos deben ser muy proactivos para prevenir.

No sucederá pasivamente. Esa misma encuesta enumeró la "promoción de prácticas de intercambio de datos" (The Economist Intelligence Unit, como se cita en Anderson, 2015) como una de las principales estrategias (muy cerca de los mandatos de

arriba hacia abajo) que esos ejecutivos consideraron exitosas en la promoción de una cultura basada en datos.

Figura 1: Respuestas a la pregunta "¿Qué estrategias han demostrado ser exitosas en la promoción de una cultura basada en datos en su organización?" de una encuesta de 530 ejecutivos seleccionados por The Economist Intelligence Unit



Fuente: Anderson, 2015.

Otro atributo a tener en cuenta al construir una cultura favorable alrededor de una base de datos de calidad es la confianza. Primero, las personas deben confiar en que los datos son confiables y precisos. En segundo lugar, deben confiar en que los datos se utilizarán con buenos resultados y no en su contra. Por ejemplo, en un hospital, "un médico temía que sus datos médicos estuvieran disponibles para el personal de la sala de emergencias, no quería que vieran sus notas en caso de que cometiera un error". Las personas tienen que superar esto y concentrarse en aumentar la calidad general de sus datos. Tercero, y esto retoma el segundo tema de esta sección, es proporcionar un amplio acceso al personal en su conjunto. Las organizaciones basadas en datos son mucho más abiertas y transparentes; y los datos están democratizados, accesibles para muchas personas dentro de la organización. "Todos los miembros de la organización deben tener acceso a la mayor cantidad de datos legalmente posible" (Patil, DJ. Manson, H. 2015). El acceso puede ser a través de informes y paneles estáticos, pero también de "acceso activo", en términos de herramientas de inteligencia comercial o incluso los datos sin procesar. Esto también implica un gran elemento de confianza.

La organización debe confiar en que los datos no caerán en una gestión abusiva, se filtrarán a los competidores o se utilizarán para impulsar batallas políticas, sino que se utilizarán de manera adecuada para promover el negocio en su conjunto. Yendo más allá, una organización basada en datos tiene un mayor potencial para impulsar la toma de decisiones más abajo en el organigrama y hacia los márgenes.

Si más colaboradores tienen acceso a los datos que necesitan; las habilidades necesarias para analizarlos e interpretarlos; y hay suficiente confianza, entonces se puede democratizar más la toma de decisiones. Por ejemplo, imagine al gerente de una tienda minorista que puede usar las herramientas de inteligencia comercial provistas para analizar las ventas de SKU en su tienda; realizar descomposición de tendencias estacionales; tener en cuenta las condiciones locales, como el clima o la construcción; pronosticar hábilmente las tendencias; y realizar pedidos de reabastecimiento para evitar los desabastecimientos y minimizar los niveles de inventario en la parte posterior de la tienda. Obviamente, muchas decisiones, especialmente las importantes o estratégicas, seguirán fluyendo hacia los niveles superiores de gestión. Sin embargo, en la mayoría de las organizaciones hay muchas decisiones, sobre todo operativas, que podrían abordarse en los márgenes, siempre que los datos sean correctos; y las habilidades y la confianza estén en su lugar. Es como nuestro sistema nervioso. La mayoría de las decisiones se envían al cerebro para su procesamiento, pero, si pisa una tachuela, se produce un reflejo espinal en el que el estímulo llega solo hasta la columna antes de regresar a los músculos para mover la pierna. El procesamiento y la toma de decisiones "locales" son suficientes para resolver el problema.

3.1.8 Amplia alfabetización de datos

Ahora bien, un concepto que va acompañado de bases de datos de calidad y una cultura de datos es la denominada "alfabetización de datos". La alfabetización de datos es la capacidad de leer, trabajar, analizar y comunicarse con datos. Es una habilidad que permite a los trabajadores de todos los niveles hacer las preguntas correctas sobre datos y sistemas; generar conocimientos; tomar decisiones; y comunicar el significado a los demás.

Por poner un ejemplo, claramente, los analistas necesitan capacitación en diseño experimental, pensamiento crítico, presentación de datos, uso de herramientas de inteligencia comercial, estadísticas, etc. Sin embargo, para que una empresa se base en datos, este conjunto de habilidades y perspectiva basada en evidencia y hechos debe ser arraigado a una escala mucho más amplia. Necesita que los gerentes y otros tomadores de decisiones también conozcan los datos, pero ¿Por qué? Aquí te brindamos algunas importantes razones:

- Los gerentes firman los cheques para comprar, instalar y mantener una nueva herramienta de inteligencia comercial o un servicio de modelado predictivo. Tienen que entender el valor que esto traerá a la organización.

- Los gerentes aprueban la interrupción del flujo de trabajo de sus equipos y sufren una productividad reducida a medida que los analistas toman clases,

se capacitan y aprenden nuevas herramientas. En resumen, reciben un golpe durante la transición, por lo que deben comprar las ganancias a largo plazo.

- Los gerentes toman las decisiones estratégicas y tácticas finales basadas en los análisis. Tienen que reconocer las fallas y luego retroceder cuando se les presenta un análisis descuidado. Deberían estar constantemente haciendo preguntas más profundas, más ricas y más indagatorias sobre los datos y esperando más de los analistas. También tienen que presentar sus hallazgos y conclusiones a los más altos directivos, al directorio o a los inversionistas. Es decir, deben comprender los matices del análisis, tener confianza en él y estar preparados para defenderlo.

Como hemos visto, dentro del marco de una estrategia de implementación de *people analytics* en donde la calidad de datos es importante, también es imprescindible considerar un plan para consolidar una cultura de datos y desplegar acciones para fortalecer las capacidades asociadas a la alfabetización de datos.

Referencias

Anderson, C. (2015). *Creating a Data-Driven Organization*. O´Reilly Media.

Patil, D. J. Manson, H. (2015). *Data Driven*. O´Reilly Media.

Tableau. *(2021). ¿Qué es una cultura de datos?* https://www.tableau.com/es-es/why-tableau/data-culture#:~:text=Una%20cultura%20de%20datos%20son,la%20identidad%20de%20una%20organizaci%C3%B3n.