Manual Imprescindible

Arquitectura e ingeniería de datos

Pilares sólidos para decisiones estratégicas

Walter E. Calcagno Lucares

ANAYA MULTIMEDIA

ndice de contenidos

Introducción y cómo usar este libro

Introducción

No se puede iniciar un libro de arquitectura e ingeniería de datos sin reflexionar sobre la gran cantidad de datos que a diario generamos en el planeta.

Según el sitio ia-latam.com, se estima que para el año 2025 estaremos generando 463 exabytes diarios de *data* en el mundo llevando la acumulación anual de estos datos a los 181 zettabytes en 2025, una cifra de orden astronómico.

Consideramos que un zettabyte se representa como un 1 seguido por 21 ceros de bytes, eso quiere decir mil trillones de bytes, ya es una cifra grande, ahora multiplica eso por 181; realmente, un número grande.

A pesar de dicha magnitud, aproximadamente el 1 % de todos estos datos generados son efectivamente analizados para tomar decisiones. En 2013, un informe de IDC y EMC Corporation estimaba que solo alrededor del 0,5 % de los datos generados en el mundo se analizaban. Desde entonces, hemos visto un crecimiento exponencial en la generación de datos y por supuesto que también hemos observado el mismo crecimiento en la capacidad de almacenamiento y análisis de datos gracias a la aparición y el perfeccionamiento del Big Data.[1]

Sin embargo, a pesar de los avances en análisis distribuido, inteligencia artificial, aprendizaje profundo y supercómputo, todavía es probable que solo una fracción de los datos generados se analice efectivamente.

Más de un 95 % de los datos del mundo se encuentran aún sin analizar, según las publicaciones anteriores. Esto evidencia una alta demanda de analistas de datos y de profesionales que sepan cómo construir soluciones analíticas robustas y escalables con las distintas herramientas que en la actualidad existen y que se presentan principalmente en las distintas nubes como SaaS (*Software as a Service*) y PaaS (*Platform as a Service*).

Seguramente, ya has leído la frase "la *data* es el nuevo petróleo", pero el petróleo en su estado más puro no sirve para nada; para darle una utilidad debo refinarlo y transformarlo en combustible o plásticos. Lo mismo pasa con la *data*: si no soy capaz de obtener *insights* o patrones en los datos, no voy a poder transformar aquello en una decisión o en una llamada a la acción.

---

1. https://ia-latam.com/2019/04/18/cuanta-data-se-genera-en-un-dia/.
   https://es.statista.com/grafico/26031/volumen-estimado-de-datos-digitales-creados-o-replicados-en-todo-el-mundo/.
   Gantz, J. y Reinsel, D. (2012). "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East". IDC iView: IDC Analyze the Future. Recuperado de https://www.cs.princeton.edu/courses/archive/spring13/cos598C/idc-the-digital-universe-in-2020.pdf.

La arquitectura de datos, en esencia, consiste en definir qué herramientas se van a montar en algún espacio físico o virtual para extraer la *data*, procesarla y obtener de dicho proceso los modelos analíticos que entregarán constantemente los *insights* necesarios para la toma de decisiones.

Gracias a una buena arquitectura de datos, se pueden desenvolver las soluciones analíticas que soportan los múltiples análisis que existen en la ciencia de datos, tanto en la disciplina conocida como inteligencia de negocios como en la analítica avanzada, donde se incluye la analítica prescriptiva, predictiva, el aprendizaje automático y la inteligencia artificial.

Un arquitecto de datos debe actuar como un director de orquesta o un maestro constructor. Debe conocer cómo organizar los distintos elementos o piezas que forman un ecosistema analítico, siendo muy riguroso con las fundaciones de su construcción y también lo suficientemente innovador para permitir que en un futuro cercano nuevas formas de analizar los datos y generar valor a la organización puedan integrarse y convivir con los elementos ya existentes.

También debe considerar que el mayor valor obtenido por los distintos análisis nacidos desde la solución analítica debe ser superior a los costes que implementar dicha solución implica. En términos financieros, siempre debe tener en cuenta el retorno de la inversión o ROI (*Return of Investment*, según sus siglas en inglés).

Una guía para decidir si vale la pena la relación coste beneficio se llama "Matriz, Valor, Complejidad", plasmada por Mertens & Van Baelen en su libro *Azure Data and AI Architect Handbook* (2023), que, sin traducirlos literalmente, lo expresan de esta forma: la matriz valor-complejidad se sostiene sobre dos ejes que van de menos a más, el eje X representa la complejidad de obtener una solución analítica escalable o que logre los resultados, y el eje Y representa el valor para la organización de dicha solución analítica. Al dividir esta figura en cuadrantes, veremos que el cuadrante superior izquierdo o también cuadrante de alto valor-baja complejidad es el que contendrá los llamados *Quick Wins* o "cosecha rápida". Por otro lado, los que se encuentren en el cuadrante inferior derecho o también cuadrante bajo valor-alta complejidad contendrá las soluciones *No Go* o no priorizables por tener un muy bajo ROI. La matriz se observa de esta forma en la figura I.1.

Antes de construir, siempre se debe planificar el paso a paso de una implementación; partiremos con un boceto general hasta la elaboración de las ingenierías de detalles. Este plan debe considerar la capacidad de escalar para manejar volúmenes de datos en constante crecimiento, la capacidad de adaptarse a diferentes tipos y estructuras de datos, y la capacidad de integrarse con diferentes tecnologías y sistemas existentes. Además, debe tener en cuenta los aspectos de seguridad y privacidad de los datos, garantizando que la información personal y comercial confidencial esté protegida de acuerdo con las leyes y regulaciones aplicables.

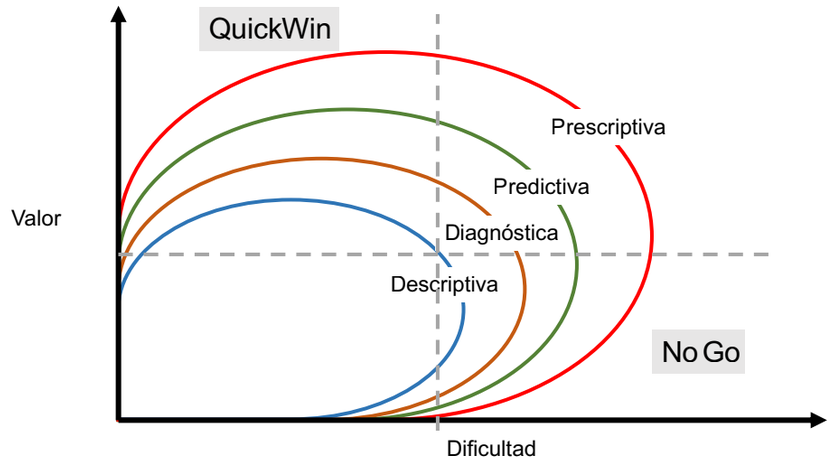

Figura I.1. Matriz valor complejidad, elaboración propia basada en Mertens & Van Baelen.

Pero quizás lo más importante de todo es que los arquitectos de datos deben ser conscientes de la necesidad de traducir los datos en información significativa. Los datos en sí mismos son solo números y palabras; el verdadero valor proviene de convertir esos datos en información que puede ser utilizada para tomar decisiones informadas. Y, para hacer eso, necesitamos entender no solo cómo almacenar y procesar los datos, sino también cómo interpretarlos y comunicar los resultados de manera efectiva.

Cómo usar este libro

Este libro está dirigido para todas las personas que tengan la responsabilidad de construir soluciones analíticas en sus organizaciones, independientemente del tamaño que estas tengan, ya sean líderes ejecutivos como gerentes, vicepresidentes, subgerentes o jefaturas de áreas que tomen decisiones basados en datos, y también está dirigido hacia analistas de datos que deseen transformarse en ingenieros o arquitectos de datos o que deseen adquirir una visión holística del ecosistema de datos que soporta las decisiones que ayudan a tomar con su trabajo. Para lograr todo lo anterior, este libro se ha estructurado de la siguiente forma.

2

Roles en la ciencia de datos

En este capítulo aprenderás:

- Cuáles son los roles en torno a la ciencia de datos.
- Cuáles son los roles en torno al gobierno de datos.
- Cuáles son los roles en torno a la ejecución de proyectos.
- Cuáles son los roles en torno al desarrollo de software.

Roles en la ciencia de datos

Ya definimos anteriormente en el capítulo 1 qué es la ciencia de datos, sus objetivos y cómo se relaciona muchísimo con la transformación digital. En este capítulo, haremos un recorrido general por todos los roles actualmente demandados desde la perspectiva de la ciencia de datos y ampliaremos un poco el espectro hacia algunos roles que también tienen mucha relación con la ciencia de los datos, pero que no necesariamente son parte de este conjunto de disciplinas.

Sin embargo, estamos obligados a trabajar directamente con ellos puesto que en una organización flexible, ya sea de cara a la implementación de un proyecto o de cara al ejercicio de una función permanente, la interacción y el trabajo en equipo son vitales para el logro de los objetivos.

Para nuestro asombro, muchos de los roles que nombraremos a lo largo de estas páginas no necesariamente están ejercidos en la actualidad por profesionales de formación universitaria de escuelas asociadas a la informática, sistemas o estadísticas, sino que, para sorpresa de muchos lectores (y mía también), dichas funciones son ejercidas por profesionales y expertos de las más distintas áreas del saber: historiadores del arte, contables, diseñadores gráficos, médicos, odontólogos, físicos y astrofísicos, ingenieros industriales, deportistas, artistas visuales y comunicadores, por nombrar algunos que he conocido en mi carrera que son parte de esa experimentada capacidad de ejercicio.

Esta diversidad de carreras de origen hace que las miradas hacia la ejecución de proyectos de implementación o ejecución de estas tareas sean muy ricas en ideas, innovación y adopción a lo largo de las organizaciones. Sin duda, podemos afirmar que, gracias a esta diversidad de carreras de origen, el avance en el desarrollo de la ciencia de datos ha tenido un crecimiento exponencial y demanda la continuidad del aprendizaje de todos estos profesionales.

Para armar una buena estrategia, es necesario conocer muy bien nuestras piezas y cuáles son sus características y limitaciones. Seguramente esta frase utilizada por muchísimos jugadores de ajedrez adquiere bastante sentido cuando se trata de un relato de cara a un tablero, aunque también me atrevo a utilizarla en este capítulo puesto que un profesional que se encuentra en sus primeros pasos en la analítica debe conocer cuáles son estos roles para desarrollar su propia estrategia de aprendizaje e incluso decidir si los caminos de aprendizaje que actualmente se encuentran presentes principalmente de cara a la autoformación son los que necesita seguir o eventualmente debería tomar otras rutas de aprendizaje.

Los roles que a continuación describiremos los agruparemos en los siguientes grupos:

- Roles y profesionales en la ciencia de datos.
- Roles en gobierno de datos.
- Roles en ejecución y seguimiento de proyectos.
- Roles en desarrollo de software.

Roles y profesionales en la ciencia de datos

En los años 80, cuando era un niño, escuché en un cortometraje de la casa de animación Walt Disney una frase atribuida a Galileo Galilei: "Las matemáticas son el lenguaje en el que Dios escribió el universo". Al cabo de unos años, descubrí que dicha frase fue adaptada artísticamente de la original que versa: "Las leyes de la naturaleza están escritas en el lenguaje de las matemáticas... los símbolos son triángulos, círculos y otras figuras geométricas, sin cuya ayuda es imposible comprender una sola palabra".

Cuando hablamos del propósito filosófico de la búsqueda de la verdad a través de los datos inevitablemente debemos ser capaces de entender la naturaleza de los mismos datos. En un simple ejercicio lógico, podríamos postular que todos los datos se pueden interpretar a través de su comportamiento y todo comportamiento obedece a patrones de conducta. Sin dudarlo, podemos afirmar que las matemáticas están presentes en la ciencia de datos y, ya que definimos que el objetivo de la ciencia de datos es la búsqueda de la verdad, podemos inferir entonces que la verdad, al parecer, está escrita matemáticamente.

¿No me crees? Bueno, hagamos un pequeño listado de algunas cosas que han podido realizar estos profesionales de la ciencia de datos:

- Han descubierto y descrito mundos completos a través de la teoría de conjuntos.
- Han podido, gracias a Bayes, entender qué sucede cuando una variable depende de otra.
- Han podido desarrollar algoritmos de predicción de comportamientos con teoría de juegos.
- Han encontrado, en el hermoso camino de la detección de patrones, las respuestas a sus preguntas de negocios.
- Han justificado la armonía de los trazos y proporciones con progresiones matemáticas como la de Fibonacci y aplicar dichas proporciones a herramientas de inteligencia artificial para hacer más fácil la vida de las personas.

Por lo tanto, dominar las matemáticas para entender y desentrañar esos patrones de conducta son una habilidad especial para todos quienes desempeñen los roles que describiremos a continuación.

Sin más preámbulo ni reflexiones, los principales roles que actualmente el mercado laboral demanda para trabajar en organizaciones que utilizan activamente la ciencia de datos como un área dentro de su estructura organizativa son los siguientes:

- Arquitectos de datos.
- Ingenieros de datos.
- Analistas de negocios.
- Analistas avanzados o científicos de datos.

Arquitectos de datos

Los arquitectos de datos son los directores de orquesta en el mundo de la analítica de datos. Su función es similar a la de un arquitecto en la construcción: debe diseñar, planificar y supervisar la infraestructura en la que se almacenarán, se accederán y se gestionarán los datos para transformarlos en productos analíticos explotables por cualquier elemento al interior de la organización e incluso fuera de ella.

Es el responsable de definir los elementos de infraestructura analítica y cómo estos elementos se correlacionan, debe validar que sus diseños no tengan brechas de seguridad o riesgos de pérdidas de información, debe definir los estándares de procesamiento y almacenamiento de datos al interior de la organización, asegurando que estos sean coherentes, de alta calidad y que sean usables por cualquier otro rol que participa en los procesos de analítica.

El arquitecto de datos juega un rol vital en la definición de cómo una organización utiliza sus datos para tomar decisiones, impulsar la eficiencia y fomentar la innovación. Su labor es esencial para garantizar que los datos sean un activo valioso y estratégico en cualquier organización.

Parece que el arquitecto de datos tiene un rol muy operativo; sin embargo, su mirada debe ser altamente estratégica, por ejemplo: adoptar una nube pública o una nube privada conlleva la ejecución presupuestaria de varios

6 El análisis predictivo

En este capítulo aprenderás:

- Qué es el análisis predictivo y sus orígenes.
- Fundamentos de aprendizaje automático e inteligencia artificial.
- Modelos de regresión lineal y análisis en series de tiempo.
- Principios de arquitectura para el aprendizaje automático e inteligencia artificial.

El análisis predictivo

Es quizás el análisis predictivo de datos la disciplina más antigua de la civilización. El ser humano desde tiempos inmemoriales ha tratado de predecir el comportamiento de distintos fenómenos naturales, el movimiento de las estrellas, la aparición de cometas, las llegadas de las lluvias y el término de estas, cuándo habría que migrar para buscar nuevas fuentes de recolección y caza y un sin número de actividades propias de la supervivencia, tanto en las épocas cuando éramos unos simples cazadores y recolectores hasta cuando domesticamos las plantas y los animales, dando origen a la civilización y las ciudades.

La responsabilidad de la predicción recaía siempre en algunos iluminados que, generalmente por gracia divina, profetizaban la ocurrencia de un hecho, aunque la evidencia acumulada a través de los años demuestra que lo que realizaban era la interpretación de patrones de algunos fenómenos; incluso, a medida que las civilizaciones antiguas fueron evolucionando, optaron por ir dejando registro de dichos patrones para poder traspasar ese conocimiento a futuras generaciones.

En otras palabras, gracias a esa capacidad de observación y deducción, pudimos aprender y luego traspasar ese conocimiento a futuras generaciones. Este conocimiento aprendido lo fuimos refinando y clasificando en disciplinas de aprendizaje, estas disciplinas las sistematizamos dando origen a las distintas ramas de la ciencia. Uno de estos conjuntos de conocimientos adquiridos fueron precisamente las matemáticas.

Imaginemos el siguiente escenario en una antigua ciudad del Oriente Medio cuando dejamos de ser cazadores recolectores para transformarnos en agricultores. En ese escenario, un pequeño agricultor llamado Adán planta un saco de trigo y cosecha, después de sus cuidados, 3 sacos de trigo; al año siguiente, planta 2 sacos de trigo y cosecha 6 sacos de trigo; al año siguiente, planta 3 sacos de trigo y cosecha 9 sacos de trigo. ¿Qué pasaría si al siguiente año Adán plantara 5 o 7 sacos de trigo?

Nuestro cerebro es capaz de identificar a través de la observación de datos de entrada y de salida una lógica o patrón de comportamiento, ese patrón de comportamiento obedece a una secuencia lógica de pasos sucesivos que interpretados matemáticamente nos permiten resolver un problema, en este caso cuántos sacos de trigo se cosecharán si Adán planta algunos.

Al adquirir el conocimiento de la secuencia lógica o fórmula, simplemente para responder algunas preguntas sobre el futuro aplicamos la fórmula deducida del paso anterior para calcular el resultado correspondiente a la aplicación de la fórmula en el problema.

Después de todo lo anterior, parece muy elemental: ordenamos los datos de entrada y salida y, a través de la observación, deduciremos el algoritmo predictor; y después de aplicar dicho algoritmo, valga la redundancia, podremos predecir el siguiente número. Por supuesto, ya que predijimos uno, podemos predecir todos los que queramos en el futuro porque detectamos ese patrón de comportamiento, como se ve en la tabla 6.1.

Tabla 6.1. Patrón de comportamiento de trigo cosechado.

Trigo plantado	Trigo cosechado
1	3
2	6
3	9
5	¿?
7	¿?

En consecuencia, podemos definir x como el valor del trigo plantado e y como el valor del trigo cosechado. Nuestra observación de los datos x e y nos lleva a deducir que:

$$y = 3x$$

Para cada valor de sacos plantados de trigo, Adán va a cosechar 3 veces ese volumen. Lo que nos lleva a concluir que para los valores 5 y 7, Adán cosechará 15 y 21 sacos respectivamente.

Parece simple, ¿no?

Egipto y el Nilo

El antiguo Egipto fue un gran ejemplo en gestión política basada en análisis predictivo de datos al gestionar la alimentación de la civilización completa (y los impuestos a recaudar), gracias a los patrones detectados en el comportamiento del río Nilo. El Nilo era la fuente de vida de Egipto, sus inundaciones anuales traían agua y suelo fértil necesario para la agricultura. Sin embargo, estas inundaciones variaban en intensidad y tiempo. Predecir estas inundaciones era vital para la planificación agrícola y la supervivencia.

Los egipcios observaron el cielo para predecir las inundaciones del Nilo. El evento clave era la aparición de la estrella Sirius, conocida como Sothis en Egipto. Esta aparición coincidía con el inicio de la temporada de inundaciones. Con base en estas observaciones, desarrollaron un calendario solar de 365 días, dividiéndolo en tres estaciones: inundación, siembra y cosecha.

Para ir corroborando lo anterior, los egipcios construyeron estructuras llamadas nilómetros a lo largo del río para medir el nivel del agua. Estos nilómetros, que eran muy parecidos a escalinatas que descendían al río, estaban marcados con inscripciones indicando niveles críticos del agua. Los sacerdotes y escribas registraban estas mediciones, las cuales eran usadas para predecir la calidad de la cosecha y determinar los impuestos. Cabe destacar que en la actualidad podemos visitar los nilómetros de Elefantina y Kom Ombo si alguna vez vamos a Egipto.

Utilizando registros acumulados durante años, los egipcios desarrollaron modelos predictivos rudimentarios. Estos modelos se basaban en correlacionar niveles pasados del Nilo con los resultados de las cosechas. Aunque no eran modelos matemáticos en el sentido moderno, representaban una forma temprana de análisis predictivo basado en datos.

Las predicciones del Nilo influían en la planificación agrícola y administrativa. Si se preveía una inundación débil, se tomaban medidas para almacenar alimentos y racionar el agua. En años de inundaciones fuertes, se preparaban para extensas siembras. Por supuesto, a pesar de todos estos esfuerzos, el sistema predictivo egipcio no estaba exento de errores. Las variaciones en el clima y factores exógenos no formaban parte de los modelos ni registros, como las guerras, las inundaciones o sequías en otras latitudes, etc.

El enfoque de Egipto en la predicción y gestión del Nilo puede verse como un precursor del análisis predictivo moderno. Su énfasis en la observación, registro e interpretación de datos es fundamental en la ciencia y la tecnología actuales.

Este legado del antiguo Egipto, a pesar de lo rudimentario de sus métodos en comparación con los actuales, demuestra que el deseo y la necesidad de comprender y predecir nuestro entorno es una búsqueda humana atemporal.

Mayas predictivos

La gestión predictiva de la civilización maya es un ejemplo fascinante de cómo una comprensión profunda de los ciclos naturales y astronómicos puede ser aplicada de manera práctica para el beneficio de una sociedad. Su habilidad para prever y planificar, a pesar de los desafíos, destaca la importancia de la observación cuidadosa y el análisis detallado en cualquier esfuerzo de planificación a largo plazo. Los mayas no solo construyeron una civilización impresionante en términos de arte y arquitectura, sino que también desarrollaron un sistema de gestión predictiva que, en muchos aspectos, era sorprendentemente moderno en su enfoque y aplicación.

9 Tipos de datos

En este capítulo aprenderás:

- Qué tipos de datos existen y cómo se clasifican.
- Tipos implícitos.
- Según su almacenamiento físico.
- Según su encriptación y compresión.

Existen diversas clasificaciones de datos dependiendo de su naturaleza, de su origen físico, de su almacenamiento y, honestamente, dependiendo de cualquier particularidad existente o por existir. A pesar de esta dificultad de establecer algunos términos de clasificación generalmente aceptados, en este capítulo trataremos las principales clasificaciones de tipos de datos existentes, lo cual nos ayudará en la formulación de estrategias de arquitectura e ingeniería de datos, valga la redundancia.

Clasificación implícita

Los datos se pueden clasificar implícitamente, es decir, en lo que representan de forma implícita, en dos grandes grupos: los categóricos y los numéricos, los que, a su vez, se dividen en discretos y continuos, como se ve en la figura 9.1.

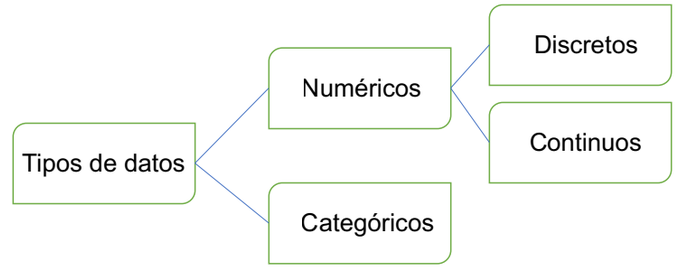

Figura 9.1. Tipos de datos implícitos (elaboración propia).

Datos numéricos

Son datos numéricos aquellos que representan y pertenecen al mundo de las matemáticas. Dentro de esta categoría, podemos distinguir dos tipos:

- **Datos continuos:** Aquellos sobre los cuales podemos aplicar todo tipo de pruebas aritméticas y estadísticas. En otras palabras, son datos continuos todos aquellos elementos que pertenecen al conjunto de los números reales e imaginarios, por ejemplo, la distancia recorrida por un vehículo, la altura de una montaña, las unidades de memoria consumidas por un programa, etc.
- **Datos discretos:** Son aquellos cuyos elementos pertenecen al conjunto de los números enteros, es decir, que no son divisibles ni fraccionables, por ejemplo, la cantidad de hijos de una persona, ya que no puede tener 2,5 hijos, o tiene 2 o tiene 3; la cantidad de vehículos que se pueden estacionar afuera de una oficina, pueden ser 3 o 4, pero no 3,25 vehículos.

Datos categóricos

Los datos categóricos son aquellos no numéricos que se pueden agrupar, contar, clasificar y, en algunos casos, jerarquizar, por ejemplo, colores, nombres, direcciones, etc. Con el fin de estudiarlos, el analista puede jerarquizar estos datos, por ejemplo, así:

1. **Jerarquía de una compañía:** Director general, director de Finanzas, jefe de área, supervisor, operario, etc.
2. **Jerarquía de tiempo:** Año, semestre, trimestre, mes, día, horas, minutos, segundos, etc.
3. **Jerarquía geográfica:** País, estado o región, provincia, municipio, pueblo, villa, calle, etc.

Los datos categóricos también se pueden utilizar para efectos de clasificación, pero no necesariamente jerarquizada, por ejemplo, color rojo para elementos peligrosos, amarillo para los leves y verde para los que cumplen una determinada norma de peligrosidad, etc.

Clasificación de tipo informática

No puedo dejar de mencionar que, desde el punto de vista informático, se clasifican los datos basados en sus propiedades binarias para lograr una mejor compilación de estos en las máquinas encargadas de su procesamiento. Entonces, es correcto decir que, en muchos casos, deben definirse antes de la creación de algún elemento recopilatorio, ya sea un arreglo, una lista o una tabla en una base de datos, etc. Principalmente distinguimos los siguientes tipos de datos en la clasificación que vemos en la tabla 9.1.

Tabla 9.1. Clasificación de tipos de datos.

Tipo dato informático	Formato en memoria o lenguaje
Números enteros	`integer`, 16, 32, 64 bits
Números reales	`float` y `double`, 32 y 64 bits
Números complejos	`complex`
Caracteres	`char`, `varchar`, `string` de distintas longitudes
Lógicos	`boolean`, `1-0`, `true/false`
Nulos	`null`, `nonetype`

Es importante destacar que cada uno de estos tipos de datos, a su vez, tiene otras clasificaciones y que, en definitiva, cada ambiente de desarrollo, sea este un *framework* de programación o de administración de bases de datos, posee una distinta gama de tipos de datos para optimizar su compilación, ya que cada tipo de datos posee un tamaño distinto. A mayor tamaño, se requiere mayor procesamiento, como se ve en la tabla 9.2.

Tabla 9.2. Tamaño de tipos de datos.

Tipo de dato	Descripción	Tamaño
`byte`	Enteros de 8 bits	1 byte
`short`	Entero corto de 16 bits	2 bytes
`integer`	Entero de 32 bits	4 bytes
`long`	Entero largo de 64 bits	8 bytes
`single/float`	Real de 32 bits	4 bytes
`double`	Real de 64 bits	8 bytes
`decimal`	Real de 128 bits	16 bytes
`boolean`	Lógico	2 bytes
`date`	Fechas o tiempo	8 bytes
`char`	Carácter de 16 bits	2 bytes
`object`	Objeto	4 bytes
`string`	Cadena de texto	Indefinidos bytes (depende del largo del texto)

A cada uno de estos elementos, además, se los puede compilar como constantes y variables, lo que, estimado lector, te llevará a imaginar una cantidad inmensa de combinaciones posibles de tipos de datos, dependiendo del objetivo para el cual fueron creados los softwares que se soportan con dichos datos.

Es por esto, la importancia para cada analista de datos, que debe comprender la esencia de lo que se analiza para no caer en sesgos de interpretación.

Según su forma de almacenamiento físico

A medida que las tecnologías han avanzado, distintos elementos se han utilizado para almacenar físicamente los datos. Podríamos remontarnos a la antigua Mesopotamia donde las tablillas de arcilla secadas al sol fueron los primeros almacenes de datos. Luego avanzamos hacia el papiro, el papel y una serie de otros soportes físicos que acumulaban inscripciones, pero que eran bastante difíciles de "leer rápidamente", es decir, solo se "compilaban" a velocidad de ojo humano.

13 Capas de arquitectura

En este capítulo aprenderás:

- Diseñar arquitecturas de datos con algunas herramientas *on-premise* y Azure.
- Las siete capas de una arquitectura de datos.
- Diseños de alto nivel.
- Diseños de detalles.

Las siete capas de arquitectura

Como lo introdujimos en el capítulo 11, las capas de arquitectura de un diseño arquitectónico base son cuatro capas principales y tres capas transversales. En este capítulo, profundizaremos en los principales desafíos que implica abordar estas capas desde el punto de vista de diseño arquitectónico de datos.

Las arquitecturas de datos son ecosistemas vivos donde trabajan una serie de personas, estas personas en general tienen un rol clave en cada etapa de la generación de valor desde los datos.

En general, nos encontraremos con ingenieros de datos, analistas de datos, científicos de datos, equipos de ciberseguridad, equipos de gobierno de datos y por supuesto a los arquitectos.

Como es de esperarse, en cada capa de arquitectura trabajarán principalmente algunos equipos de generación de valor; por lo tanto, es muy probable que las herramientas que aquí se mencionen puedan ir cambiando dependiendo de la plataforma y el alcance de los equipos que ahí trabajan.

La composición de equipos y su correspondencia en las capas de arquitectura se puede observar en la tabla 13.1.

Tabla 13.1. Composición de equipos y correspondencia en capas.

Grupo	Capa	Disciplinas
Principal	Origen e ingesta	Ingeniería de datos
	Procesamiento y almacenamiento	
	Servicios	
Transversal	Consumo	Analistas y científicos de datos
	Gobierno	Gobierno de datos
	Seguridad	Ciberseguridad
	Monitoreo	Arquitectura de datos

Como se aprecia, los ingenieros de datos son los que más herramientas y protagonismo tienen en este ordenamiento de capas. También están los analistas y científicos de datos y, por último, el resto de los equipos transversales.

A continuación, explicaré cada una de las capas de datos y las consideraciones principales para el diseño de arquitecturas en cada capa, como una lista de requisitos y habilidades o documentación que hay que tener en cuenta antes de definir herramientas que ocupen sus lugares en cada capa.

Capa de origen e ingesta de datos

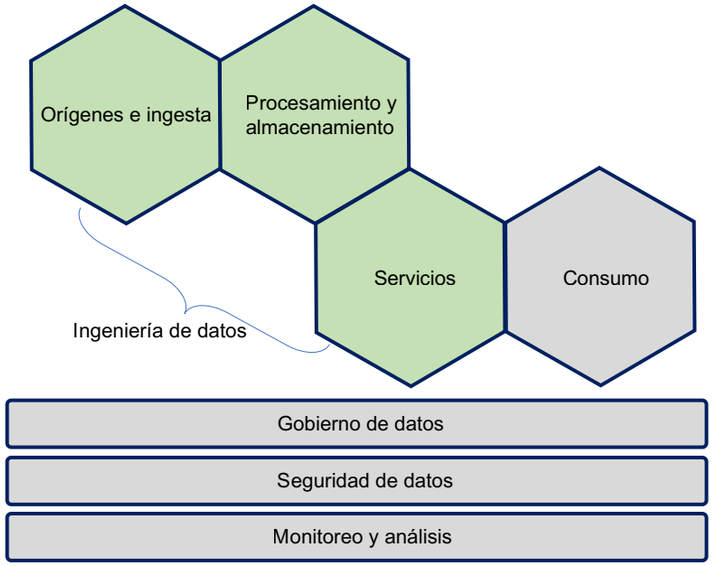

Figura 13.1. La ingeniería de datos dentro de la arquitectura de referencia.

Esta capa es la primera de lo que llamaremos ingeniería de datos, en la figura 13.1. Se caracteriza porque su objetivo es definir dos aspectos fundamentales del proceso de captura de datos: identificar el origen y definir las estrategias de captura.

Orígenes de datos

Bases de datos relacionales y NoSQL

En capítulos anteriores, ya hablamos en extenso de qué se tratan estas bases de datos, así que en esta ocasión solo enfatizaré en los requerimientos para trabajar con estos orígenes.

Para acceder a estos orígenes, se requiere una dirección de servidor, que puede ser un número IPv4 en redes privadas o una dirección IPv6 para entornos más extensos y modernos. En el caso de distribuciones en la nube, se accede mediante URL. Cada tipo de dirección refleja el entorno en el que la base de datos opera, ya sea en un entorno local (IPv4), uno más amplio y posiblemente internacional (IPv6) o en la nube, donde la accesibilidad y la escalabilidad son claves.

Estas diferencias son cruciales para entender cómo se conecta y accede a los datos, y tienen implicaciones directas en la seguridad, la latencia y la gestión general de la base de datos.

Se pueden usar tres tipos de usuarios para consultar los datos. Evidentemente, la mejor cuenta va a depender de las herramientas y servicios que se utilicen para la extracción. Estos tipos de usuario son: cuentas de servicio, usuario/contraseña de base de datos o red y usuarios con autenticación MFA:

- **Cuentas de servicio:** Estas cuentas están diseñadas para procesos automatizados y aplicaciones, no para individuos. Ofrecen control y seguridad, ya que sus permisos pueden ser restringidos específicamente para las tareas que deben realizar.
- **Usuario/contraseña de base de datos o red:** El método tradicional de acceso, donde cada usuario tiene un nombre de usuario y contraseña. Este método es simple, pero puede ser menos seguro si no se manejan adecuadamente las credenciales.
- **Usuarios con MFA (autenticación multifactor):** Ofrecen un nivel adicional de seguridad. Junto con el nombre de usuario y la contraseña, se requiere un segundo factor, como un código enviado a un teléfono móvil. Esto es ideal para proteger contra accesos no autorizados, especialmente en entornos donde la seguridad de los datos es crítica.

Por último, para acceder a estos orígenes de datos, necesitamos conocer el esquema y el catálogo de datos. Este requisito es clave al momento de hacer consultas hacia el origen, ya vimos en capítulos anteriores que por lo general vamos a encontrar estas bases de datos con un modelado entidad-relación y debemos transformar ese esquema hacia un modelado dimensional cuando hablamos de analítica; en consecuencia, tener la documentación a mano para escribir buenas consultas es un requisito clave y forma parte de la documentación de inicio que debemos solicitar cuando vamos a extraer desde una base de datos.

Archivos, ficheros u objetos binarios

Esta categoría abarca una amplia gama de formatos, desde documentos de texto y hojas de cálculo hasta archivos multimedia como imágenes y vídeos.

Para trabajar con estos datos, hay que tener en cuenta los siguientes elementos: ubicación, tipo de archivo, tamaño y encriptación. También hay que saber si para la analítica necesitamos el contenido del fichero u objeto binario o solamente la metadata, ya que eso cambia absolutamente la perspectiva de la ingesta.

16 Ingeniería de datos con Microsoft Fabric

En este capítulo aprenderás:

- Qué es Microsoft Fabric.
- Cómo se ajusta a una arquitectura general de datos.
- Componentes de ingesta en *batch* y *streaming* de Fabric.
- Componentes de almacenamiento.
- Componentes de la capa de servicios Spark de Fabric.
- Capa de consumo de Fabric.

Qué es Microsoft Fabric

De acuerdo con la documentación oficial de Microsoft, "Fabric es una solución de análisis todo en uno para empresas que abarca todo, desde el movimiento de datos hasta la ciencia de datos, el análisis en tiempo real y la inteligencia empresarial. Ofrece un conjunto completo de servicios que incluye un lago de datos, ingeniería de datos e integración de datos, todo en un solo lugar.

Con Fabric, no es necesario agrupar diferentes servicios de varios proveedores. En su lugar, puede disfrutar de un producto muy integrado, de un extremo a otro y fácil de usar diseñado para simplificar las necesidades de análisis".

Desde mi perspectiva, Microsoft Fabric, en la figura 16.1, es una solución modular que permite integrar en un solo ecosistema las distintas capas de arquitectura analítica escalable para organizaciones independientemente del tamaño que esta tenga. Está pensado para que organizaciones pequeñas que pensaban que montar una nube analítica era algo demasiado sofisticado puedan comenzar desde lo más sencillo y escalar rápidamente hacia ecosistemas más complejos.

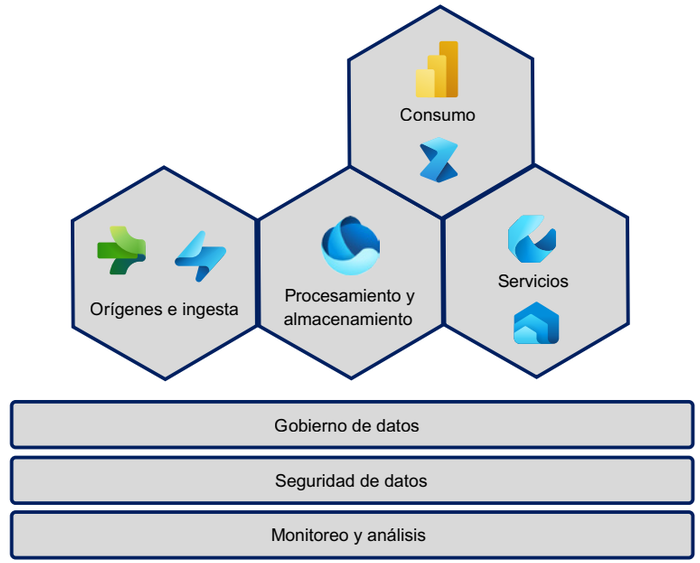

Figura 16.1. Arquitectura de Microsoft Fabric en un estándar de arquitectura tradicional.

Si se observa con detención, veremos que los distintos componentes que forman Microsoft Fabric se corresponden en las distintas capas de una arquitectura tradicional y de ellos hablaremos en las siguientes páginas.

Fabric viene a convertirse en una PaaS que combina distintos elementos ya existentes en Azure y que los fusiona para evitar pasos innecesarios de importación o conexión; principalmente, hereda de Synapse los motores Spark y SQL Serverless, también hereda Azure Data Factory, además de incorporar dos elementos originales, como son Data Activator y Copilot.

Para ir entendiendo completamente esta *suite*, la revisaremos capa a capa y veremos las bondades que nos ofrece.

Por supuesto, como ha sido la tónica de este manual, iré reforzando la explicación con ejemplos paso a paso perfectamente replicables.

Primeros pasos en Fabric

Para entrar a Microsoft Fabric, escribiremos la URL: *https://fabric.microsoft.com*.

Luego observaremos que la interfaz es muy similar a la interfaz del servicio de Power BI; abajo, a la izquierda, nos ofrecerá la interfaz donde queramos trabajar para lograr nuestro cometido analítico.

La figura 16.2 muestra esta interfaz y, para efectos del relato, iremos trabajando desde la capa de orígenes e ingesta, donde para nuestra fortuna la herramienta que nos ofrecerá es Data Factory. Algo ya hemos visto de ella, así que será muy sencillo entender cómo opera dentro de Fabric.

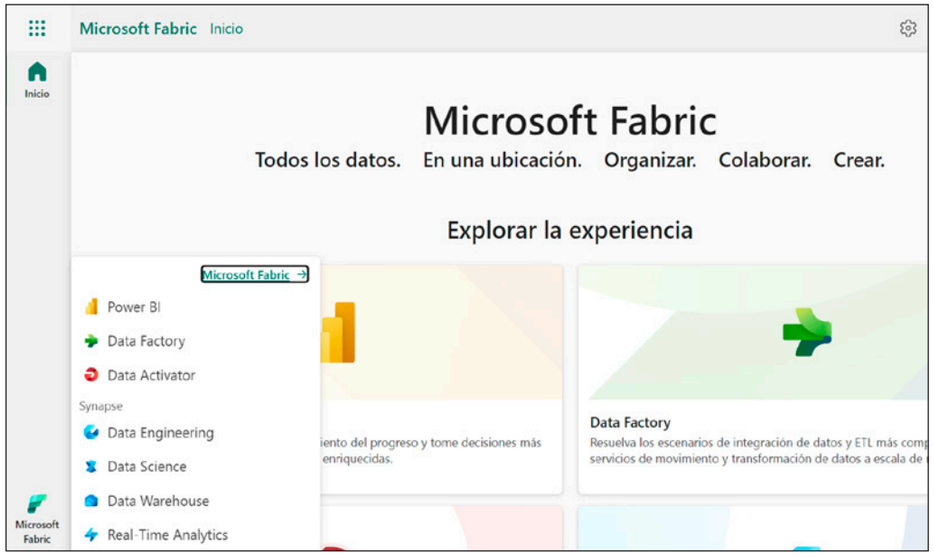

Figura 16.2. Bienvenida de Microsoft Fabric.

Capa de orígenes e ingesta

Fabric Data Factory

Es la herencia desde Azure que se integra en esta plataforma y está programado para el uso versátil de dos grandes herramientas para realizar importaciones de tipo *batch*. Estas herramientas son Flujo de datos Gen2 y Canalización de datos, las cuales pasaremos a revisar en detalle a continuación.

Flujo de datos Gen2

Los flujos de datos se crean utilizando la misma experiencia de Power Query, que está disponible hoy en día en Excel, Power BI, Power Platform, aplicaciones Dynamics 365 Insights, entre otras.

Power Query es una herramienta que presenta una interfaz de muy bajo código y que bajo el lenguaje M (proviene del inglés *Mashup*) permite de una forma muy versátil desarrollar tareas de ingesta y almacenamiento de datos en la misma plataforma. A diferencia de su antecesor, los flujos de datos de segunda generación no solo permiten consumir los datos en Power BI, sino también escribirlos en cualquier artefacto de almacenamiento dentro de Fabric.

La interfaz de los flujos de datos está pensada para analistas sin formación de código, es decir, usuarios de negocios que requieren algunas tareas de ingesta para su autoservicio. Eso permite una alta velocidad en la curva de aprendizaje y en el corto plazo tener implementadas varias tareas de ETL escalables horizontalmente.

Los flujos de datos permiten realizar combinaciones y anexiones de tablas, incorporar parámetros y crear funciones personalizadas, además de efectuar tareas de agregaciones, limpieza de datos, transformaciones personalizadas y mucho más desde una interfaz bastante amigable muy visual y fácil de usar.

Sigamos con el ejemplo para entender mejor cómo opera un flujo de datos.

Al seleccionar la interfaz de Data Factory, veremos que se nos ofrecen las áreas de trabajo. Para el libro, ya he creado una llamada `WS_Libro_AID_2024`, como se observa en la figura 16.3.

Nuestra primera acción para traer datos utilizando los flujos de datos de segunda generación será la creación de un flujo de datos. Para ello, haremos clic en el menú Nuevo y seleccionaremos el botón Flujo de datos Gen2, como se indica en la figura 16.4.

Podemos observar que hay muchos más elementos para trabajar; veremos varios de ellos en este libro.

Figura 16.3. Área de trabajo de Data Factory sin elementos.

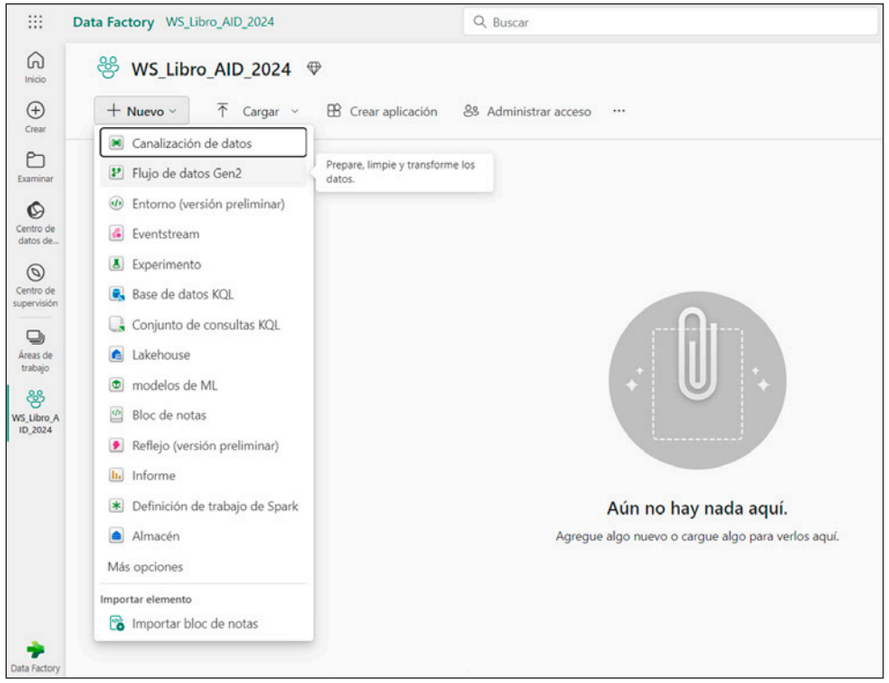

Figura 16.4. Menú de selección de nuevo flujo de datos.

Una vez que nuestra interfaz de flujos de datos se encuentra lista para comenzar a trabajar se verá como en la figura 16.5.

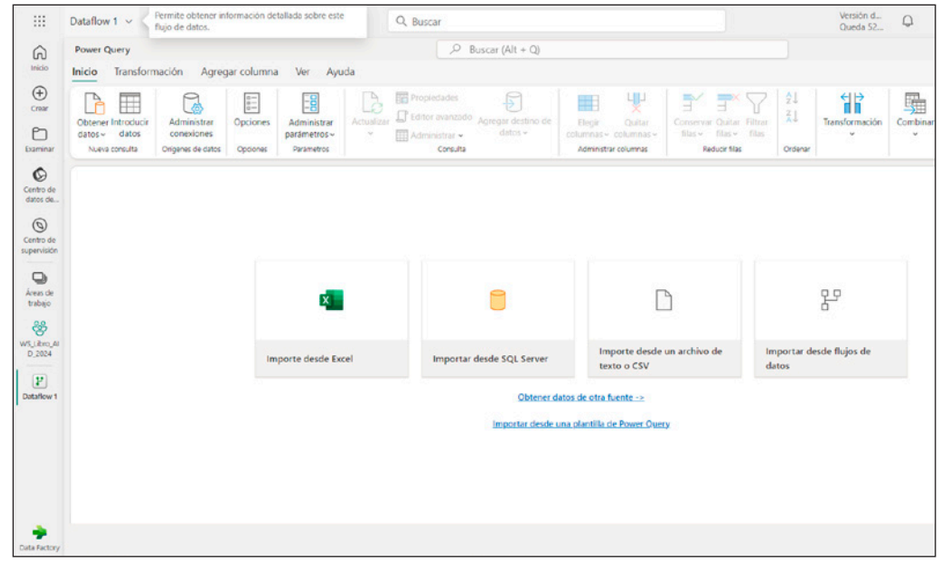

Figura 16.5. Interfaz de los flujos de datos en Fabric.

La interfaz es muy similar a la interfaz de Power Query en Power BI Desktop, por lo que debería ser muy natural para quien está acostumbrado a dicha herramienta poder utilizar esta interfaz.

A continuación, iremos a buscar nuestro *dataset* de cambio climático. La URL la conocemos y no tenemos que modificar nada, la vuelvo a poner de nuevo para evitar retroceder páginas:

```
https://raw.githubusercontent.com/MinCiencia/Datos-CambioClimatico/main/output/
temperatura_dmc/1950/1950_temperatura_dmc.csv
```

En el flujo de datos, seleccionaremos el botón Obtener datos y buscaremos la opción API web, como se ve en la figura 16.6.

Seleccionada dicha opción, escribiremos la URL en la casilla habilitada para esos efectos, como se muestra en la figura 16.7.

Hacemos clic en el botón Siguiente y veremos cómo se nos muestra la vista previa de los datos que estamos capturando desde el Ministerio de Ciencias, como se ve en la figura 16.8.

Arquitectura e ingeniería de datos

Manual Imprescindible

Esta obra proporciona una introducción accesible y completa a los conceptos clave, las técnicas y las mejores prácticas en el campo de la arquitectura y la ingeniería de datos, sin la necesidad de conocimientos previos en programación o estadísticas.

Aborda secuencialmente una descripción general de los conceptos clave en la arquitectura de datos, incluidas las definiciones esenciales hasta la descripción de los sistemas de gestión de datos, los modelos de datos, el almacenamiento y la integración de datos. Se exploran las diferencias entre las bases de datos relacionales y no relacionales, así como las ventajas y desventajas de cada enfoque. También se abordan las consideraciones de seguridad y privacidad en la arquitectura de datos, y se proporcionan pautas para garantizar la protección de la información confidencial. Luego se adentra en la ingeniería de datos, que se centra en la ingesta de datos, así como en la limpieza, el enriquecimiento y la validación de datos.

Ya seas un gerente, un analista de negocios, un consultor o simplemente alguien interesado en aprender más sobre cómo los datos pueden impulsar el éxito empresarial, este libro te proporcionará las habilidades y el conocimiento necesarios para navegar con confianza en el complejo mundo de la arquitectura y la ingeniería de datos.

ISBN 978-84-415-4970-8
2311386
9 788441 549708

ANAYA MULTIMEDIA

www.anayamultimedia.es