Antes de transformar los datos, debemos entenderlos

Data Understanding: dejar de buscar datos a ciegas y comenzar a construir conocimiento sobre nuestros datos

En la reflexión anterior planteaba una idea:

> **Antes del dashboard, debemos entender la empresa.**

Ese es precisamente el propósito de **Business Understanding**: comprender qué quiere lograr la organización, cuáles son sus objetivos, qué decisiones necesita tomar y cómo sabremos si estamos avanzando en la dirección correcta.

Si realizamos bien esa primera etapa, algo cambia radicalmente cuando llegamos a **Data Understanding**:

Ya no vamos a buscar datos a ciegas.

Ahora sabemos qué estamos buscando.

Y, sobre todo:

sabemos para qué lo necesitamos.

Esa diferencia puede transformar por completo un proyecto de datos.

---

1. De navegar entre silos a buscar datos con propósito

Durante muchos proyectos podemos caer en una situación que seguramente otros profesionales de datos reconocerán.

Comenzamos a navegar.

Una base de datos.

Después otra.

Un Excel.

Una aplicación.

Un repositorio.

Una tabla histórica.

Otro archivo que alguien ha mantenido durante años.

Seguimos explorando con la esperanza de encontrar información que nos permita construir el reporte o dashboard solicitado.

Pero quizá el problema no era que tuviéramos pocos datos.

El problema era que **todavía no sabíamos exactamente qué estábamos buscando**.

Cuando no conocemos suficientemente el objetivo del negocio, terminamos preguntándonos:

> **“¿Qué puedo hacer con los datos que encontré?”**

Después de Business Understanding podemos invertir la pregunta:

> **“¿Qué datos necesito para responder las preguntas de negocio que ya definimos?”**

La diferencia parece pequeña.

Pero cambia completamente la investigación.

---

2. Business Understanding nos dio el mapa. Data Understanding comienza la exploración

Imaginemos que durante Business Understanding establecimos que necesitamos comprender, por ejemplo:

crecimiento,

rentabilidad,

clientes,

productos,

canales,

países,

evolución temporal

y posición en determinados mercados.

Ahora Data Understanding no comienza diciendo:

“Veamos qué datos tenemos.”

Comienza preguntando:

“¿Dónde están los datos que necesitamos?”

¿En qué sistemas están?

¿Quién los genera?

¿Quién los conoce?

¿Con qué frecuencia se actualizan?

¿Qué granularidad tienen?

¿Qué período histórico existe?

¿Qué unidad de medida utilizan?

¿Qué tan completos están?

¿Podemos relacionarlos?

¿Son adecuados para el objetivo que definimos?

Ahora existe **propósito en la búsqueda**.

---

3. Encontrar un dato no significa comprenderlo

Imaginemos que encontramos una columna denominada:

VENTAS

Excelente.

¿O no?

Todavía debemos preguntar:

¿Venta facturada? ¿Venta cobrada? ¿Venta bruta?

¿Venta neta?

¿Incluye impuestos?

¿Incluye devoluciones?

¿En qué moneda está expresada?

¿A qué fecha corresponde?

¿Cuál es su granularidad?

Encontrar el campo no es suficiente.

Debemos comprender qué significa.

Y aquí aparece una conexión extraordinariamente importante entre las dos primeras etapas de CRISP-DM:

> **Business Understanding nos ayuda a definir qué significa el concepto para el negocio.**
>
> **Data Understanding nos permite investigar cómo ese concepto está representado en nuestros datos.**

Y algunas veces descubriremos que ambas cosas **no coinciden**.

Eso también es un hallazgo.

---

4. NULL no significa automáticamente cero

Este es uno de mis ejemplos favoritos para explicar la importancia de esta etapa.

Imaginemos un equipo que toma mediciones de temperatura:

23 | 24 | NULL | 25 | 27

Desde una perspectiva exclusivamente técnica alguien podría decir:

> “Tenemos un valor faltante. Cambiemos NULL por 0.”

Podemos hacerlo.

Pero antes existe una pregunta mucho más importante:

¿Qué significa NULL en este proceso?

¿El dispositivo no transmitió la medición?

¿Estaba apagado?

¿Falló? ¿Ese día no correspondía medir?

¿Se perdió el dato? ¿O realmente se registraron cero grados?

No lo sabemos.

Y precisamente esa es la enseñanza.

NULL no significa automáticamente cero.

Si sustituimos NULL por cero sin comprender el proceso que generó ese dato, podemos cambiar completamente su significado.

Entonces quizá ya no estamos limpiando los datos.

Estamos alterando la realidad que pretendemos analizar.

---

5. Antes de sumar, debemos entender la unidad

Otro ejemplo clásico.

Tenemos información de varios países.

Queremos conocer el total.

Sumamos.

¿Problema resuelto?

No necesariamente.

¿Qué sucede si una fuente está expresada en quetzales y otra en dólares?

¿Qué significa el resultado?

La operación matemática puede ejecutarse correctamente y el resultado, aun así, **no tener sentido para el negocio**.

Antes debemos comprender:

unidad de medida;

moneda;

escala;

formato;

granularidad;

periodicidad;

significado de la variable.

Antes de agregar, tenemos que saber si los datos son comparables.

Porque sumar correctamente lo que no debería sumarse sigue produciendo un resultado incorrecto.

---

6. El gran riesgo: transformar antes de comprender

Encontramos duplicados.

Los eliminamos.

Encontramos NULL.

Los sustituimos.

Encontramos valores extremos.

Los corregimos.

Encontramos formatos diferentes.

Los normalizamos.

Parece eficiente.

Pero probablemente estamos avanzando demasiado rápido.

Primero debemos preguntar:

¿Por qué existen los duplicados?

Quizá son errores.

Pero también podrían representar transacciones diferentes.

¿Por qué existe el NULL?

Podría ser un problema de calidad.

Pero podría representar algo perfectamente válido en el proceso.

¿Por qué encontramos un valor extremo?

Podría ser un error.

O podría ser precisamente el comportamiento extraordinario que nuestro análisis necesita descubrir.

Por eso considero importante establecer una frontera:

Data Understanding no consiste todavía en arreglar los datos.

Consiste en **conocerlos, describirlos, explorarlos, cuestionarlos y evaluar si pueden responder al objetivo del negocio**.

Después decidiremos cómo prepararlos.

---

7. La calidad de datos tiene un propósito

Esta etapa también nos permite comenzar a evaluar calidad.

Pero incluso aquí deberíamos evitar clasificar superficialmente un dataset como:

“bueno” o “malo”.

Una pregunta más útil sería:

> **¿Estos datos tienen la calidad necesaria para el uso que queremos darles?**

Un conjunto de datos podría ser suficiente para un análisis descriptivo y no ser adecuado para otra finalidad.

Una variable con valores faltantes podría tener poca importancia para una pregunta y ser crítica para otra.

Una fuente mensual podría ser suficiente para una decisión estratégica y completamente insuficiente para una necesidad operativa diaria.

Por eso podemos comenzar a estudiar dimensiones como:

- completitud;
- consistencia;
- validez;
- unicidad;
- exactitud, cuando tengamos una referencia contra la cual contrastarla;
- temporalidad;
- integridad entre fuentes;
- coherencia de formatos;
- coherencia de unidades.

Pero esas métricas solo adquieren sentido cuando recordamos:

¿Para qué necesitábamos estos datos?

---

8. Data Understanding también es investigación

Aquí comenzamos a comportarnos menos como constructores de dashboards y más como investigadores.

Preguntamos.

Exploramos.

Perfilamos.

Comparamos.

Analizamos distribuciones.

Revisamos frecuencias.

Identificamos valores faltantes.

Observamos duplicados.

Encontramos valores inesperados.

Contrastamos fuentes.

Buscamos relaciones.

Y, especialmente:

preguntamos por aquello que no comprendemos.

Si una categoría solamente existe en un país:

¿por qué?

Si una variable dejó de registrarse a partir de una fecha:

¿por qué?

Si encontramos diferentes identificadores para aparentemente un mismo cliente:

¿por qué?

Si una fuente se actualiza diariamente y otra mensualmente:

¿qué implica para nuestro análisis?

Cada anomalía puede ser un error.

Pero también puede revelar algo sobre **cómo funciona realmente nuestra organización**.

Y aquí descubrimos otra característica importante de CRISP-DM:

> **Lo que aprendemos de nuestros datos también puede hacernos regresar y comprender mejor el negocio.**

---

9. Una organización Data Driven debe conocer su paisaje de información

Si aspiramos a convertirnos en una organización verdaderamente **Data Driven**, Data Understanding puede dejar un legado mucho mayor que comprender los archivos utilizados por un único proyecto.

Esta etapa puede comenzar a construir progresivamente un mapa:

¿Qué datos tenemos y dónde viven?

Podemos empezar a documentar una cadena como:

Objetivo de negocio → Dato necesario → Fuente → Sistema → Responsable → Periodicidad → Granularidad → Calidad → Uso

Eso cambia profundamente el conocimiento organizacional.

Dejamos de depender de frases como:

> “Creo que Finanzas tiene ese Excel.”

o:

> “Pregunta a determinada persona; esa persona sabe dónde está.”

Y comenzamos a construir una organización que sabe **dónde están sus datos y qué representan**.

---

10. Aquí comienza nuestro inventario de fuentes

Mientras investigamos para nuestro proyecto podemos comenzar a registrar las fuentes que vamos descubriendo.

CRM.

ERP.

Core transaccional.

Archivos.

Aplicaciones.

Sistemas externos.

Bases departamentales.

Fuentes públicas.

APIs.

O cualquier otra fuente existente en nuestra realidad.

Pero no simplemente hacemos una lista.

Documentamos contexto:

qué contiene;

quién la genera;

quién responde por ella;

qué periodicidad tiene;

qué granularidad utiliza;

qué período histórico contiene;

qué calidad hemos observado;

para qué puede ser útil.

El proyecto comienza así a contribuir con algo que puede sobrevivir al proyecto:

un inventario progresivo de los activos de datos de la organización.

---

11. Y comenzamos a descubrir los silos

Podemos encontrar algo particularmente interesante.

El concepto **Cliente** aparece en:

CRM,

facturación,

créditos,

tarjetas,

canales digitales

y muchos otros sistemas.

Entonces surge una pregunta aparentemente sencilla:

¿Estamos hablando del mismo cliente?

¿Cómo se identifica en cada sistema?

¿Existe una llave compartida?

¿Tenemos múltiples identificadores?

¿Qué granularidad utiliza cada fuente?

¿Quién posee la definición?

¿Qué sistema debemos considerar como referencia para determinado atributo?

De manera progresiva empezamos a descubrir cómo los silos podrían relacionarse.

Ya no estamos aprendiendo solamente sobre un dataset.

Estamos aprendiendo sobre el ecosistema de información de la empresa.

---

12. Metadata: entender el dato más allá de su valor

Supongamos que encontramos:

SALDO = 10,000

El valor por sí solo no nos dice suficiente.

Necesitamos contexto.

¿Saldo contable?

¿Saldo disponible?

¿Saldo promedio?

¿En qué moneda?

¿De qué fecha?

¿A qué producto corresponde?

¿De qué sistema proviene?

¿Quién puede validar su definición?

Ese contexto comienza a construir nuestra **metadata**.

Podemos explicarlo de una forma sencilla:

> **El dato nos dice el valor.**
>
> **La metadata nos ayuda a entender qué significa.**

Y todavía podemos avanzar un paso más.

---

13. Del inventario al linaje

Ahora preguntamos:

¿De dónde viene este dato?

¿Qué sistema lo originó?

¿Qué procesos recorrió?

¿Fue transformado?

¿Dónde quedó almacenado?

¿En qué modelo fue utilizado?

¿En qué reporte o indicador terminó?

Conceptualmente comenzamos a construir un recorrido:

Sistema de origen → Integración → Transformación → Modelo → KPI → Dashboard

Esto comienza a proporcionarnos **trazabilidad y linaje**.

Y cuando mañana alguien diga:

> “Este indicador parece incorrecto.”

ya no empezaremos necesariamente nuestra investigación desde cero.

Podremos comenzar a recorrer el camino del dato hacia atrás.

---

14. Empezamos también a descubrir conceptos comunes

Mientras investigamos distintas fuentes podemos empezar a encontrar conceptos que atraviesan toda la organización:

Cliente

Producto

Tiempo

Geografía

Canal

Moneda

Estos conceptos pueden aparecer repetidamente en diferentes procesos.

Y esto comienza a perfilar discusiones futuras sobre integración, arquitectura, modelos de datos o un bus de datos.

Pero aquí mantendría algo importante:

Todavía estamos comprendiendo.

No necesitamos adelantarnos a construir la arquitectura definitiva.

Primero descubrimos qué existe.

Cómo se relaciona.

Qué definiciones son comunes.

Dónde existen contradicciones.

Y qué debería estandarizarse posteriormente.

La arquitectura debería construirse sobre ese entendimiento, no antes.

---

15. La IA vuelve a potenciar nuestras habilidades

Aquí continuamos con el concepto que introdujimos durante Business Understanding:

Potenciar mis habilidades con IA.

Podemos pedir a la IA que explore los datos desde diferentes perspectivas.

Como Analista de Datos

¿Qué patrones y anomalías deberíamos investigar?

Como especialista de Data Quality

¿Qué dimensiones de calidad deberíamos evaluar?

Como Data Steward

¿Qué definiciones, metadata, responsables y reglas necesitan validación?

Como Científico de Datos

¿Qué variables presentan riesgos para análisis posteriores?

Como Auditor

¿Qué situaciones podrían producir interpretaciones incorrectas?

Como experto del negocio

¿Qué valores aparentemente anómalos podrían tener una explicación operativa?

Cada rol puede ayudarnos a descubrir preguntas distintas.

Pero mantenemos exactamente el mismo principio que utilizamos en Business Understanding:

> **La IA no sustituye a quien conoce los datos. Nos ayuda a llegar con mejores preguntas.**

La IA propone.

La IA encuentra patrones.

La IA ayuda a cuestionar.

El humano valida el significado y decide.

Eso es potenciar nuestras habilidades.

---

16. El entregable: nuestro Documento de Entendimiento de los Datos

Todo este aprendizaje no debería quedarse en nuestra cabeza.

Ni exclusivamente en un chat.

Ni perdido dentro del proyecto.

El resultado debería convertirse en un nuevo activo de conocimiento:

Documento de Entendimiento de los Datos

Podría consolidar:

Datos requeridos

¿Qué información determinamos que necesitamos a partir del objetivo de negocio?

Datos encontrados

¿Qué fuentes contienen esa información?

Definiciones

¿Qué significa cada variable crítica?

Estructura

¿Cuáles son sus granularidades, identificadores y relaciones?

Unidades

¿Qué monedas, escalas o medidas encontramos?

Calidad

¿Qué problemas o particularidades identificamos?

Metadata

¿Qué contexto necesitamos preservar para interpretar correctamente los datos?

Linaje inicial

¿De dónde provienen los datos y qué recorrido realizan?

Responsables

¿Quién puede validar cada fuente o definición?

Brechas

¿Qué datos necesitamos pero actualmente no existen?

Riesgos

¿Qué situaciones podrían afectar los análisis posteriores?

Preguntas pendientes

¿Qué todavía necesitamos entender?

---

17. El documento no es estático

Este punto es fundamental.

Nuestro Documento de Entendimiento de los Datos **no representa una verdad inmutable**.

Aparecen nuevas fuentes.

Cambian sistemas.

Se modifican procesos.

Descubrimos errores.

Cambian definiciones.

Mejora la calidad.

Aparecen nuevas necesidades.

Entonces el documento evoluciona.

Podemos entenderlo así:

> **“Esto representa lo que hemos investigado, entendido y validado hasta este momento.”**

Una nueva iteración puede enriquecerlo.

Un nuevo equipo puede cuestionarlo.

Un experto puede corregirlo.

Un proyecto futuro puede ampliar el linaje.

Y progresivamente nuestra organización construye algo extraordinariamente valioso:

memoria institucional sobre sus datos.

---

18. El siguiente integrante ya no comienza desde cero

Imaginemos que mañana llega un nuevo analista.

Un científico de datos. Un arquitecto.

Un consultor.O simplemente comenzamos una nueva iteración del mismo proyecto.

Ya tenemos dos piezas fundamentales.

🧩 Business Understanding

Sabemos **qué intenta lograr la organización y por qué**.

🧩 Data Understanding

Sabemos **qué datos necesitamos, qué encontramos, qué significan y qué sabemos sobre su calidad y procedencia**.

La nueva persona no debería comenzar preguntando:

> “¿Dónde están todos los archivos?”

Puede comenzar leyendo el conocimiento que la organización ya construyó.

Entender. Cuestionar.

Validar.

Y agregar una nueva pieza.

Ese es el valor de gestionar conocimiento.

El proyecto termina.

Lo que aprendimos sobre nuestros datos permanece.

---

19. Una organización Data Driven no es simplemente una organización con muchos datos

Tampoco es necesariamente la que tiene más dashboards.

Ni la que compra más tecnología.

Ni la que utiliza más IA.

Para mí, una organización comienza a desarrollar verdaderamente una cultura Data Driven cuando progresivamente puede responder:

¿Qué datos tenemos?

¿Dónde están?

¿Qué significan?

¿De dónde vienen?

¿Qué calidad tienen para el uso que queremos darles?

¿Cómo se relacionan?

¿Quién responde por sus definiciones?

¿Qué información todavía no tenemos?

Y finalmente:

> **¿Podemos confiar en ellos para tomar esta decisión?**

En ese momento dejamos de simplemente almacenar datos.

Comenzamos a gestionarlos como un activo de la organización.

Antes de transformar, debemos comprender

Si tuviera que resumir **Data Understanding** en algunas ideas serían estas:

Business Understanding nos dijo qué buscar.

Data Understanding nos permite descubrir qué tenemos.

El inventario nos dice dónde está.

La metadata nos ayuda a entender qué significa.

La calidad nos ayuda a determinar si sirve para nuestro propósito.

El linaje comienza a mostrarnos de dónde viene.

La gobernanza empieza a clarificar quién responde por ello.

La documentación permite que lo aprendido permanezca.

Y la IA puede potenciar nuestra capacidad para investigar todo lo anterior.

Por eso:

> **Antes de reemplazar un NULL, debemos entender qué significa.**
>
> **Antes de eliminar un duplicado, debemos comprender por qué existe.**
>
> **Antes de sumar, debemos validar la unidad.**
>
> **Antes de integrar fuentes, debemos entender qué representa cada una.**
>
> **Antes de transformar los datos, debemos entenderlos.**

Porque al terminar Data Understanding no solamente deberíamos conocer mejor los datos de **este proyecto**.

Deberíamos haber conseguido algo mucho más importante:

que nuestra organización conozca un poco mejor sus propios datos que cuando comenzamos.

Y que la próxima persona, el próximo proyecto o la próxima iteración **no tenga que volver a descubrir desde cero lo que nosotros ya aprendimos.**