Cada transformación debe tener una razón… y dejar conocimiento

Data Preparation: de transformar datos a construir capacidad organizacional

En las reflexiones anteriores construimos dos principios:

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

Después:

> **Antes de transformar los datos, debemos entenderlos.**

Ahora llegamos a la tercera etapa de CRISP-DM: **Data Preparation**.

Ya conocemos el objetivo del negocio.

Sabemos qué preguntas necesitamos responder.

Identificamos los datos necesarios.

Conocemos sus fuentes, significado, unidades, granularidad, calidad y principales problemas.

Ahora sí estamos preparados para intervenirlos.

Es momento de agregar valor a los datos.

Pero hacerlo bajo un principio fundamental:

> **No transformamos porque podemos. Transformamos porque sabemos por qué debemos hacerlo.**

---

1. Ahora sí podemos transformar

Durante Data Understanding identificamos problemas.

Encontramos datos faltantes.

Duplicados.

Diferencias de formato.

Categorías inconsistentes.

Monedas diferentes.

Fechas inválidas.

Valores extremos.

Diferentes granularidades.

Problemas de integración.

Pero deliberadamente evitamos correr demasiado rápido a “corregirlos”.

Primero preguntamos:

¿Qué significa este NULL?

¿Por qué existe este duplicado?

¿Esta aparente anomalía es realmente un error?

¿En qué unidad está expresado este valor?

¿Estos dos campos representan realmente lo mismo?

Ahora, con ese conocimiento, Data Preparation nos permite actuar.

CRISP-DM contempla en esta etapa actividades como **seleccionar, limpiar, construir, integrar y dar formato a los datos**, con el objetivo de obtener el conjunto que utilizarán las etapas posteriores. Además, estas actividades pueden ser iterativas y no necesariamente ocurren una sola vez en un orden rígido.

---

2. Cada transformación comienza con el objetivo del negocio

Este principio debería acompañarnos durante toda la etapa.

No seleccionamos determinada información porque:

> “Tal vez pueda servir.”

La seleccionamos porque:

> **“La necesitamos para responder esta pregunta del negocio.”**

No creamos una nueva variable porque técnicamente podemos hacerlo.

La creamos porque tiene utilidad para nuestro propósito analítico.

No integramos todas las fuentes disponibles porque queremos crear el dataset más grande posible.

Integramos aquellas que aportan información necesaria para el objetivo.

La selección de los datos y la justificación para incluirlos o excluirlos forman parte precisamente de esta etapa de CRISP-DM.

Por eso:

Cada transformación debe tener una razón de negocio, analítica o de calidad que podamos explicar.

---

3. Calidad de datos: del diagnóstico a la acción

Durante Data Understanding diagnosticamos.

Ahora comenzamos a intervenir.

Podemos trabajar sobre aspectos como:

- completitud;
- consistencia;
- validez;
- unicidad;
- formatos;
- unidades;
- relaciones entre fuentes;
- categorías;
- fechas;
- identificadores;
- valores faltantes;
- anomalías previamente investigadas.

Pero nuevamente:

NULL no significa automáticamente cero.

Duplicado no significa automáticamente eliminar.

Outlier no significa automáticamente error.

Ahora transformamos porque anteriormente comprendimos el contexto.

La preparación busca elevar la calidad de los datos al nivel necesario para el análisis que posteriormente realizaremos.

---

4. No destruyamos nuestra referencia original

Aquí incorporaría una práctica que considero estratégica:

Preservar el dato base.

Podemos explicarlo utilizando el concepto de capas.

🥉 Bronze: preservar el origen

Conservamos nuestra referencia de los datos recibidos.

Necesitamos poder responder mañana:

¿Qué recibimos originalmente?

Porque puede suceder que una regla cambie.

Que descubramos que una transformación fue incorrecta.

Que una nueva necesidad requiera volver al origen.

O que necesitemos reproducir exactamente un resultado anterior.

El dato original es nuestra referencia.

🥈 Silver: enriquecer con propósito

Sobre esa base comenzamos a agregar calidad y valor.

Normalizamos.

Estandarizamos.

Validamos formatos.

Aplicamos reglas de calidad.

Convertimos unidades.

Integramos.

Creamos atributos.

Anonimizamos cuando corresponde.

Y aplicamos las transformaciones requeridas por nuestro objetivo.

Empezamos a construir una relación fundamental:

Dato original → Regla aplicada → Dato transformado

Esta lógica no pertenece formalmente a CRISP-DM. Es una forma práctica de complementar Data Preparation con buenas prácticas de arquitectura, trazabilidad y gobierno.

---

5. Estandarizar significa construir un lenguaje común

Supongamos que encontramos:

F
Fem
Femenino
femenino

Si el negocio confirma que representan el mismo concepto, podemos establecer una regla común.

Lo mismo podría ocurrir con:

Guatemala | GT | GUA

o con productos, canales, regiones y muchas otras variables.

Esta transformación aparentemente pequeña tiene un significado mayor:

estamos construyendo un lenguaje común para los datos.

Y ese lenguaje común mejora nuestra capacidad para comparar, integrar, analizar y posteriormente automatizar.

---

6. Transformar no significa necesariamente sobrescribir

Durante nuestras transformaciones podemos mantener, cuando resulte apropiado:

`pais_original` → `pais_estandarizado`

`monto_original` → `monto_moneda_comun`

`fecha_original` → `fecha_validada`

Esto facilita algo fundamental:

comparar el antes y el después.

¿Cuántos registros cambiaron?

¿Qué regla los afectó?

¿Cuáles permanecieron igual?

¿Qué problemas no fueron resueltos?

¿Qué resultado produjo la transformación?

Así, la preparación deja de ser una caja negra.

Se vuelve visible y explicable.

---

7. También comenzamos a construir nuevos datos

Data Preparation no significa solamente corregir.

También significa **enriquecer**.

Podemos obtener nuevas variables a partir de información existente cuando el objetivo lo requiera.

Por ejemplo:

Ingreso - costo → margen

Fecha de nacimiento → edad

Fecha de transacción → periodo

Transacciones → frecuencia

CRISP-DM contempla la construcción de nuevos atributos derivados como parte de Data Preparation.

Pero nuevamente debemos poder responder:

¿Por qué creamos esta variable?

Y documentar cómo fue construida.

---

8. Integrar fuentes también requiere conocimiento

Después de recorrer nuestros silos, quizá necesitemos integrar información.

Pero dos fuentes correctas pueden producir un resultado incorrecto si las relacionamos incorrectamente.

Por eso preguntamos:

¿Cuál es la llave?

¿Cuál es la granularidad?

¿La relación es uno a uno?

¿Uno a muchos?

¿Podemos duplicar registros accidentalmente?

¿Podemos perder información?

¿La agregación mantiene el significado del dato?

La integración de información procedente de múltiples tablas o registros es también una actividad contemplada en esta etapa de CRISP-DM.

---

9. El verdadero Cliente 360 no consiste en juntar todo

Este principio también resulta válido para una estrategia de Cliente 360.

Un Cliente 360 no debería convertirse simplemente en:

> **“Todas las columnas que tenemos sobre el cliente.”**

Debería preguntarse:

> **“¿Qué conocimiento del cliente necesitamos integrar para responder correctamente al objetivo del negocio?”**

Cross-selling puede requerir una visión.

Riesgo podría requerir otra.

Retención podría necesitar otra.

La preparación de datos debe estar guiada por **propósito**, no por acumulación.

---

10. Y mientras transformamos datos, enriquecemos el conocimiento

Aquí aparece una parte que considero trascendental.

Data Preparation no debería enriquecer solamente el dataset.

También debería enriquecer nuestros activos de conocimiento.

El proyecto comienza a fortalecer progresivamente:

📖 Diccionario de Datos

Qué significa cada dato original y transformado.

🗂️ Inventario de Datos

Qué fuentes y conjuntos de información tenemos, qué contienen y para qué los utilizamos.

🏷️ Metadata

Qué contexto necesitamos para interpretar correctamente nuestros activos.

🔗 Linaje

De dónde viene el dato, qué transformaciones recibe y dónde termina.

🧭 Mapa de dependencias o bus de datos

Cómo se relacionan nuestros conjuntos de información y qué productos dependen de ellos.

📋 Registro de Transformaciones

Qué cambiamos, por qué, cómo y bajo qué regla.

📘 Documento de Preparación de Datos

La consolidación del conocimiento generado durante toda esta etapa.

Estos elementos complementan CRISP-DM desde perspectivas de arquitectura, operación, gobierno y gestión del conocimiento.

---

11. El linaje ahora puede contarnos la historia completa

Durante Data Understanding preguntábamos:

¿De dónde proviene este dato?

Ahora agregamos:

¿Qué pasó con él después?

Podemos visualizar conceptualmente:

Sistema de origen

↓

Dato base

↓

Transformación

↓

Dato estandarizado

↓

Integración

↓

Producto analítico

↓

Dashboard / Modelo / Decisión

Si mañana detectamos un problema en un indicador, ya no necesitamos investigar completamente a ciegas.

Podemos recorrer el camino hacia atrás.

Esto es trazabilidad.

12. El mapa de datos también nos ayuda cuando algo falla

Supongamos que nuestro dashboard gerencial depende de tres conjuntos de información:

A + C + D

Si tenemos documentadas esas dependencias, sabemos qué fuente alimenta cada elemento.

Y si además conocemos nuestros procesos y su comportamiento histórico, podemos gestionar una interrupción con mayor evidencia.

¿Qué proceso falló?

¿Qué otros procesos dependen de él?

¿Qué productos de información fueron afectados?

¿Qué debemos reconstruir?

¿Qué secuencia debe recuperarse?

¿Qué antecedentes tenemos sobre los tiempos del proceso?

Esto no significa que podamos garantizar exactamente cuánto tardará la recuperación. Cada incidente puede tener condiciones diferentes.

Pero sí cambia la calidad de nuestra conversación gerencial.

De:

> “Ya estamos trabajando y ya casi está.”

a:

> **“Conocemos qué proceso fue afectado, qué productos dependen de él, qué debemos recuperar y tenemos antecedentes que nos permiten administrar y comunicar mejor la recuperación.”**

Ese es un ejemplo concreto de cómo el conocimiento de nuestros datos puede convertirse también en **capacidad operativa**.

---

13. El tiempo de transformación también puede generar conocimiento

También podemos comenzar a aprender del comportamiento de nuestros procesos.

¿Cómo evolucionan los tiempos?

¿Qué cambió entre versiones?

¿Aumentó el volumen?

¿Cambió una fuente?

¿Se incorporó una nueva regla?

¿Existe algún cuello de botella?

¿La lógica necesita revisión?

¿El equipo necesita fortalecer alguna capacidad?

El objetivo no debería ser utilizar esta información como mecanismo de vigilancia individual.

Debería utilizarse para:

entender los procesos y mejorarlos.

La autoría y los responsables sí deben quedar documentados para preservar conocimiento, facilitar mantenimiento y saber dónde obtener contexto.

Pero si descubrimos una práctica mejorable, la oportunidad es:

mejorar, estandarizar, automatizar o capacitar.

No buscar culpables.

Eso es verdadera mejora continua.

---

14. Documentar una transformación significa poder reproducirla

Imaginemos nuevamente una conversión monetaria.

No necesitamos saber solamente:

1,000 → 130

Necesitamos saber:

¿Qué fórmula se utilizó?

¿Qué tasa?

¿Qué fuente proporcionó esa tasa?

¿Qué fecha correspondía?

¿Qué versión de la regla estaba vigente?

¿Qué transformación produjo el resultado?

Cuando podemos responder esas preguntas conseguimos algo fundamental:

reproducibilidad.

Si mañana debemos regresar y ejecutar el proceso bajo las mismas condiciones, tenemos la posibilidad de hacerlo.

Si encontramos un error en determinada fórmula, podemos identificar qué resultados debemos revisar.

Si mejoramos la regla, podemos documentar qué cambió.

Por eso:

> **Documentar no es burocracia. Es construir la capacidad de explicar, repetir, corregir y mejorar.**

---

15. Las mejores prácticas necesitan una estrategia

Aquí está uno de los elementos que considero más importantes.

Podemos implementar:

Bronze.

Silver.

Metadata.

Linaje.

Diccionario.

Inventario.

Bus de datos.

Versionamiento.

Registro de transformaciones.

Pero utilizar conceptos de arquitectura no nos convierte automáticamente en una organización Data Driven.

La pregunta debe ser:

¿Para qué estamos haciendo todo esto?

Preservamos el origen **para recuperar y auditar**.

Documentamos **para reproducir**.

Estandarizamos **para comparar**.

Registramos linaje **para rastrear**.

Construimos metadata **para comprender**.

Mapeamos dependencias **para anticipar impactos**.

Observamos tiempos **para aprender y mejorar**.

Registramos autoría y decisiones **para preservar conocimiento**.

Usamos IA **para potenciar nuestra capacidad**.

Las buenas prácticas generan verdadero valor cuando responden a una estrategia.

---

16. Y aquí la IA puede potenciar extraordinariamente nuestras habilidades

Data Preparation es probablemente una de las etapas donde más visible puede hacerse la utilidad de la IA.

Podemos utilizar diferentes perspectivas.

Como especialista de Data Quality

¿Qué problemas necesitan tratamiento?

Como Data Engineer

¿Qué transformaciones podrían estandarizar e integrar los datos de acuerdo con nuestras reglas?

Como Data Steward

¿Las transformaciones preservan las definiciones acordadas?

Como Científico de Datos

¿Podría alguna transformación alterar de manera importante el análisis posterior?

Como especialista de privacidad

¿Qué información requiere especial consideración conforme al propósito y las reglas aplicables?

Como Auditor

¿Qué cambió entre el dataset original y el transformado?

La IA también puede ayudarnos a explicar lógica existente, proponer transformaciones, documentarlas inicialmente y comparar resultados.

Pero aquí necesitamos reforzar más que nunca nuestro principio:

La IA propone y acelera. El humano comprende, valida y decide.

Porque ahora la IA ya no nos ayuda únicamente a formular preguntas.

Ahora estamos modificando datos.

---

17. La IA puede escribir la transformación. Eso no significa que debamos ejecutarla ciegamente.

Esta distinción es fundamental.

Podemos solicitar:

> “Genera la transformación para estandarizar estas categorías.”

O:

> “Propón la lógica para convertir estas monedas.”

O:

> “Identifica duplicados.”

La IA puede reducir muchísimo el esfuerzo técnico.

Pero el profesional mantiene la responsabilidad de:

revisar la lógica;

validar la regla;

comprobar el resultado;

comparar antes y después;

documentar la transformación.

Podemos resumirlo así:

> **La IA acelera la transformación. El humano responde por su significado.**

---

18. No basta con transformar: debemos volver a medir

Al finalizar el proceso no deberíamos decir simplemente:

> **“Ya limpiamos los datos.”**

Debemos volver a evaluar.

Si Data Understanding nos proporcionó una línea base, Data Preparation debería permitirnos comparar.

Antes → Transformación → Después

¿Qué problemas fueron resueltos?

¿Qué indicadores de calidad mejoraron?

¿Qué problemas permanecen?

¿Qué riesgos residuales existen?

¿Qué cambios inesperados aparecieron?

¿Qué requiere todavía validación humana?

Así aparece otro ciclo:

Transformar → Medir → Comparar → Validar → Documentar

Y, si es necesario:

Mejorar → volver a transformar.

---

19. El objetivo tampoco es perseguir una perfección abstracta

No siempre necesitamos corregir absolutamente todo.

Tampoco necesitamos utilizar todas las fuentes.

Ni integrar todos los sistemas.

La pregunta es:

> **¿Los datos tienen ahora la calidad y preparación necesarias para responder responsablemente al objetivo del proyecto?**

La calidad debe mantenerse conectada al propósito.

Un proyecto puede necesitar determinado nivel de granularidad, temporalidad o completitud y otro proyecto puede necesitar uno diferente.

El negocio sigue siendo nuestra brújula.

---

20. El primer gran entregable: datos preparados y trazables

Después de este trabajo debemos contar con datos que, según las necesidades del proyecto, hayan sido:

seleccionados;

limpiados;

estandarizados;

normalizados cuando corresponda;

integrados;

enriquecidos;

convertidos a unidades comparables;

protegidos conforme al contexto aplicable;

estructurados al nivel requerido;

y validados nuevamente.

Pero no entregamos solamente un dataset.

Entregamos un dataset cuyo proceso de construcción podemos explicar.

---

21. El segundo gran entregable: Documento de Preparación de Datos

Este es el elemento que conecta nuevamente CRISP-DM con la gestión del conocimiento.

Nuestro documento final debería consolidar, al menos:

Objetivo del proyecto

Por qué estos datos fueron preparados.

Datos seleccionados

Qué incluimos, qué excluimos y por qué.

Línea base de calidad

Qué encontramos antes de transformar.

Reglas de calidad

Cómo tratamos los problemas validados.

Estandarizaciones y normalizaciones

Qué unificamos y bajo qué definición.

Transformaciones

Qué hicimos y cuál fue la justificación.

Variables derivadas

Qué construimos y para qué.

Integraciones

Qué fuentes relacionamos y bajo qué reglas.

Datos originales y transformados

Cómo mantenemos trazabilidad entre ambos cuando corresponda.

Metadata y linaje

Qué debemos conocer para interpretar y rastrear el dato.

Dependencias

Qué procesos o productos utilizan esa información.

Validación

Qué cambió entre el antes y el después.

Riesgos residuales

Qué problemas todavía permanecen.

Decisiones y versiones

Qué regla se encuentra vigente y qué cambió.

Responsabilidad y conocimiento

Qué personas o funciones pueden proporcionar contexto para mantener y mejorar el proceso.

El dataset representa **el resultado técnico**.

El documento representa **el conocimiento organizacional generado para obtenerlo**.

Necesitamos ambos.

---

22. Este conocimiento tampoco es estático

Cambiarán los datos.

Cambiarán las fuentes.

Cambiará el negocio.

Mejorará nuestra lógica.

Encontraremos errores.

Aparecerán nuevas tecnologías.

Cambiaremos procesos.

Entonces también deberán evolucionar:

el diccionario;

el inventario;

la metadata;

el linaje;

las dependencias;

las transformaciones;

y nuestro Documento de Preparación.

No documentamos para congelar la manera de trabajar.

Documentamos precisamente para tener una base desde la cual mejorarla.

---

23. Nuestra tercera pieza del LEGO ya está construida

Nuestra organización ahora comienza a tener una memoria metodológica:

🧩 Business Understanding

Sabemos qué quiere lograr la organización.

↓

🧩 Data Understanding

Sabemos qué datos necesitamos, dónde están y qué significan.

↓

🧩 Data Preparation

Sabemos cómo los transformamos, por qué los transformamos y cómo podemos reproducir y mejorar ese proceso.

Ahora, cuando se incorpora una nueva persona o comienza una nueva iteración, no tiene por qué comenzar desde cero.

Puede conocer:

qué decidió el negocio;

qué aprendimos de los datos;

qué problemas encontramos;

qué transformaciones realizamos;

qué reglas utilizamos;

qué riesgos quedaron pendientes;

y por qué tomamos determinadas decisiones.

Puede **leer, comprender, cuestionar y mejorar**.

Y agregar la siguiente pieza.

Data Preparation no termina cuando el dato está limpio

Para mí, esta es la gran conclusión de esta etapa:

> **Un dato preparado sin documentación puede resolver la necesidad de hoy.**
>
> **Un dato preparado, trazable y documentado puede convertirse en conocimiento reutilizable para resolver también las necesidades de mañana.**

Por eso Data Preparation debería dejarnos algo mayor que:

“Tenemos datos limpios.”

Debería permitirnos afirmar:

“Tenemos datos preparados, validados y trazables, y tenemos el conocimiento documentado para explicar cómo llegamos hasta ellos.”

Porque:

Transformar agrega valor al dato.

Documentar preserva el conocimiento.

El linaje permite rastrear.

La metadata permite comprender.

La reproducibilidad permite verificar.

Las buenas prácticas permiten estandarizar.

La IA permite potenciar nuestras capacidades.

El criterio humano permite validar.

Y la mejora continua permite que todo lo anterior evolucione.

Por eso mantendría como principio central de esta etapa:

Cada transformación debe tener una razón… y dejar conocimiento.

Porque una organización verdaderamente Data Driven no debería limitarse a transformar datos.

Debería aprender de cada transformación que realiza.