# Calidad de Datos


**Yubar Daniel Marín Benjumea**

Estadístico

M.Sc en Ingeniería de Sistemas


Febrero 23 de 2022


SERVICIO GEOLÓGICO COLOMBIANO

Facultad de Minas
Sede Medellín

UNIVERSIDAD NACIONAL DE COLOMBIA

# ¿Por qué es importante la calidad de datos?

Un estudio realizado por IBM en 2016 estimo como sólo en Estados Unidos, se generan pérdidas de 3 billones de dólares por baja calidad de los conjuntos de datos (Redman, 2016) .

Un estudio realizado en *The KPMG 2017 Global CEO Outlook* que estimo como el 56% de los directores ejecutivos en las empresas se encuentran preocupados por la integridad de los datos que están usando en la toma de decisiones (KPMG, 2017).

Se ha demostrado como es común encontrar porcentajes entre 10% y un 50% de datos erróneos en los conjuntos de datos (Johnson, Leitch y Neter, 1981; Laudon, 1986; Morey, 1982) .

Data Warehouse necesarios para implementar procesos de BDA se consume entre el 30% y el 80% del tiempo de desarrollo y presupuesto buscando mejora de la calidad en los datos

# ¿Qué genera la falta de calidad en los datos?

- Sistemas antiguos.
- Mal diseño de la interface de recolección de datos.
- Falta de capacitación.
- Cambios en los procesos.
- Reglas comerciales obsoletas.
- Falta de gobierno de datos.
- Falta de cultura del dato.
- Entre otros

La calidad de datos es un proceso dentro del gobierno de datos no todo el gobierno de datos en si.

Las investigaciones indican que muchos problemas de calidad de datos son causadas pro falta de compromiso de la organización con los datos de alta calidad

# ¿Por qué garantizar la calidad en los datos?

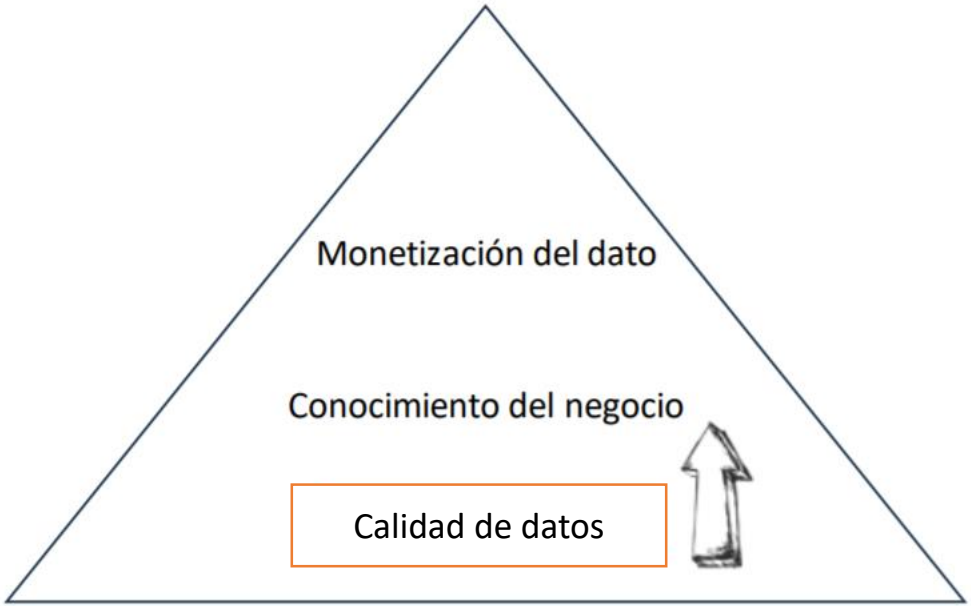
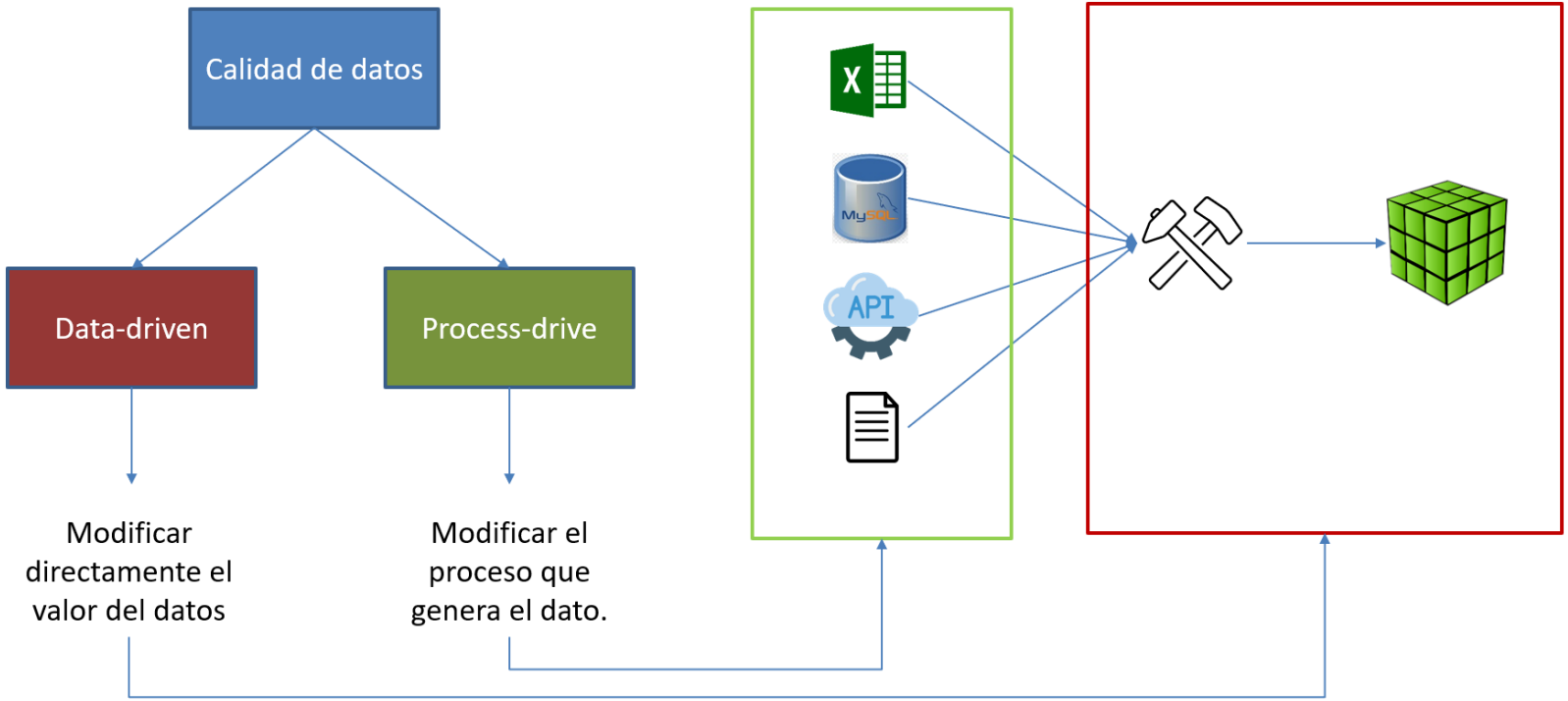



Figura 1. Gestión de los datos en una organización

# Calidad de datos

La calidad de datos se refiere tanto a las **características asociadas con los datos de alta calidad como a los procesos utilizados para medir o mejorar la calidad de los datos** (Mosley, Brackett, Earley y Henderson, 2010),

# Estrategias de calidad de datos



# Estrategias de calidad de datos

**Data driven**

Muy útil a la hora de implementar procesos de:

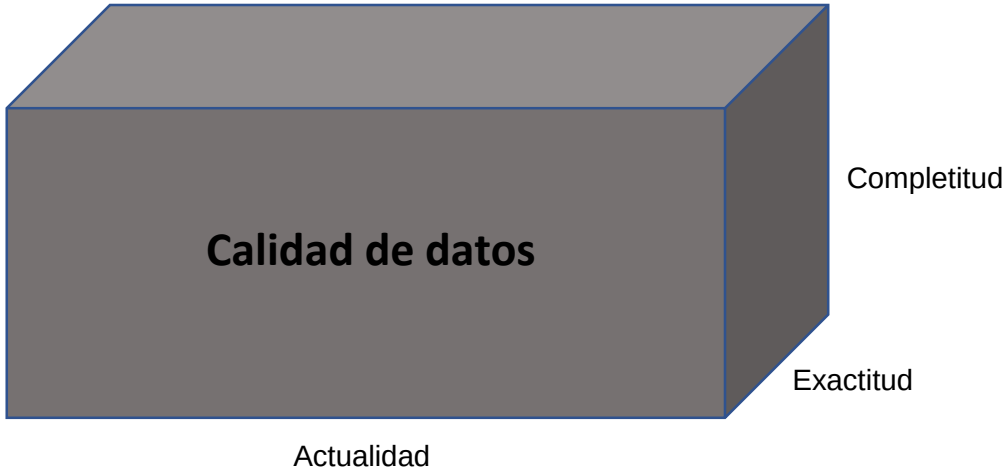
- Gestión de datos maestros.
- Data Warehouse.
- Data Mart.
- Data Lake

**Process driven**

Muy útil a la hora de implementar procesos de:

- Estructurar procesos de datos desde cero.
- Arreglar problemas de datos desde la captura.
- Solucionar problemas de calidad de datos de raíz.

# Dimensiones de calidad de datos



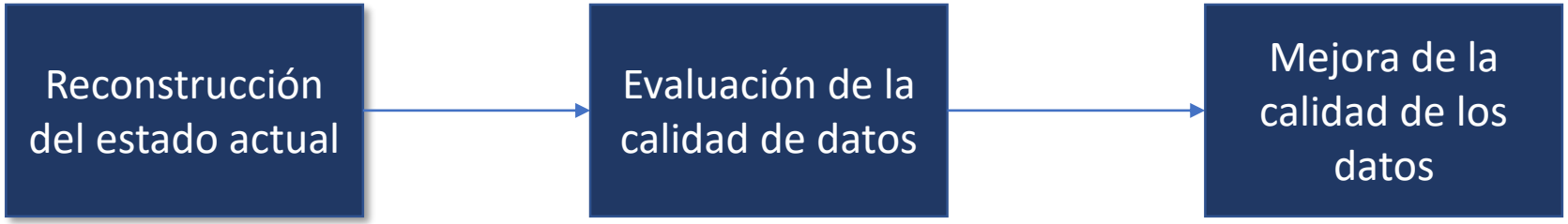

**En la literatura**

Se han mapeado 26 dimensiones (Wand y Wang, 1996) . Incluso la metodologías como COLDQ se usan 35. Lo que puede ser difícil de manejar.

**Nuestra propuesta**

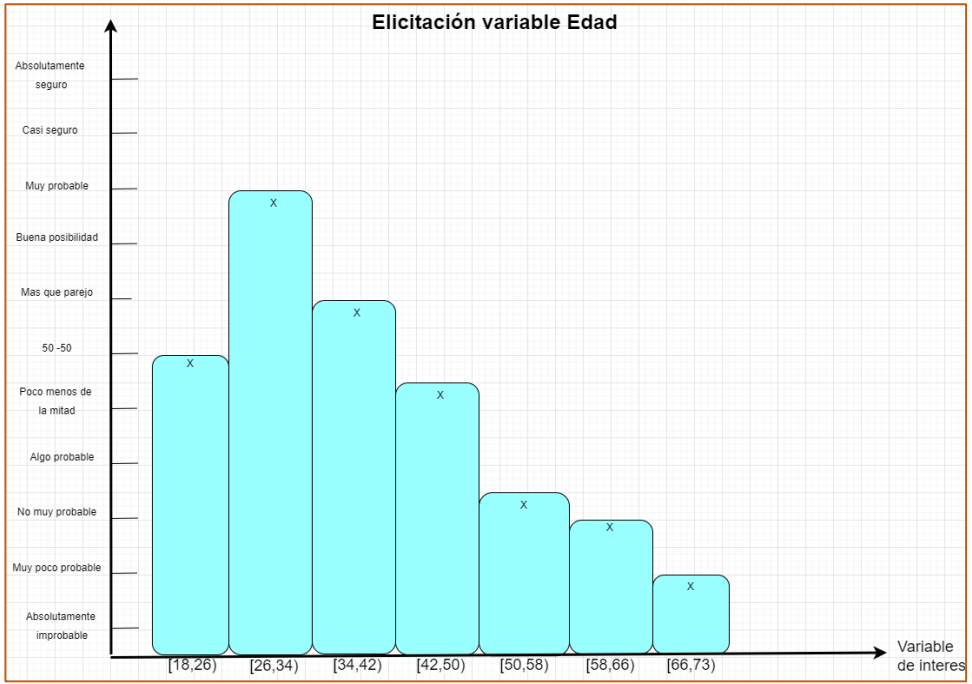
Exactitud, actualidad, completitud, consistencia y accesibilidad

# Etapas en la calidad de datos



# Reconstrucción del estado actual

- Máximos.
- Mínimos.
- Distribuciones teóricas.
- Porcentaje admisible de nulos.
- Periodos de actualización.
- Formas de acceder.
- Reglas de negocio.
- Niveles teóricos.
- Entre otros metadatos.



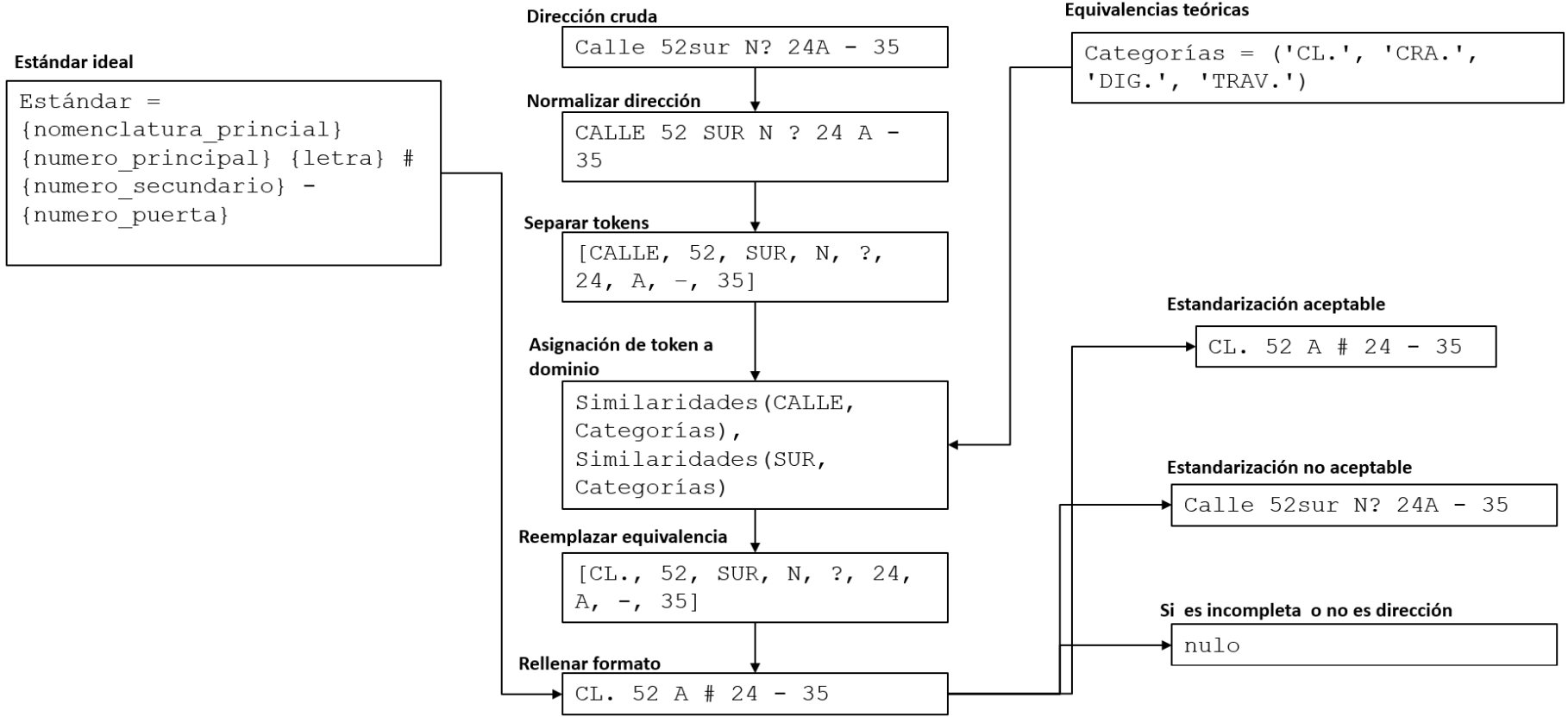
# Evaluación de la calidad o perfilamiento de datos

Definir métricas para evaluar las dimensiones

$$\%\text{Exactitud} = \left( \text{I} * 0.3 + \left( \left( \sum_{i=1}^{n} \frac{NR_i}{R_i} \right) / \text{TR} \right) * 0.7 \right) * 100\%$$

$$\%\text{Exactitud} = \left( 0.1 * \frac{\text{FMA}}{n} + 0.1 * \frac{\text{FMI}}{n} + \text{I} * 0.3 + \left( \left( \sum_{i=1}^{n} \frac{NR_i}{R_i} \right) / \text{TR} \right) * 0.5 \right) * 100\%$$

# Mejora de la calidad en los datos



SERVICIO GEOLÓGICO COLOMBIANO

Facultad de Minas
Sede Medellín

UNIVERSIDAD NACIONAL DE COLOMBIA

# GRACIAS