
# Calidad de los datos

Curso de Ingeniería de Características

https://mcd-unison.github.io/ing-caract/`

# Introducción

En la era de la información, los datos son el activo más valioso. Sin embargo, su valor depende crucialmente de su calidad.

La calidad de los datos se refiere a su precisión, integridad y confiabilidad.

# ¿Por qué importa la calidad de datos?

La mala calidad de datos puede tener consecuencias graves en cualquier ámbito:

- **Análisis inexacto:** Resultados sesgados y conclusiones erróneas.
- **Toma de decisiones deficiente:** Decisiones basadas en información incorrecta.
- **Pérdida de tiempo y recursos:** Corrección de errores y limpieza de datos.
- **Pérdida de confianza:** En los datos y en los análisis.

# Dimensiones de la calidad de datos

- **Integridad:** ¿Los datos están completos y no faltan valores?
- **Validez:** ¿Los datos cumplen con las reglas y restricciones definidas?
- **Unicidad:** ¿Hay datos duplicados?
- **Consistencia:** ¿Los datos son coherentes entre diferentes fuentes y representaciones?
- **Actualidad:** ¿Los datos están actualizados y reflejan la realidad actual?
- **Precisión:** ¿Los datos son correctos y libres de errores?

# Reglas de calidad de datos

- **Reglas de formato:** Definen el formato correcto de los datos (ej. fecha, número, texto).
- **Reglas de rango:** Definen los valores válidos para un campo (ej. edad entre 0 y 120).
- **Reglas de unicidad:** Aseguran que los datos sean únicos (ej. no haya duplicados).
- **Reglas de dependencia:** Definen las relaciones entre los datos (ej. la ciudad debe estar asociada a un país).
- **Reglas de integridad:** Verifican la completitud de los datos (ej. no debe haber campos vacíos).

# Proceso de DQA

1. Explorar y comprender los datos y sus fuentes
2. Establecer reglas de calidad
3. Integrar las reglas en un proceso automatizado de verificación
4. Monitoreo continuo de calidad de los datos

Qalyptus Data Quality Dashboard

## Overall DQ score

97

## Total rows processed

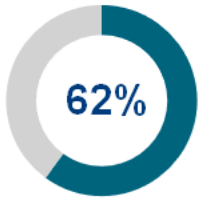
895,623

## Failed rows

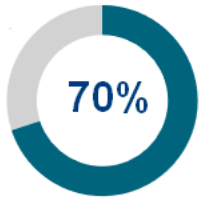
26,314

**Completeness**

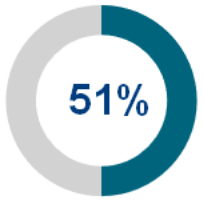

**Timeliness**




**Validity**

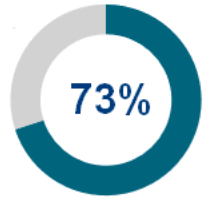



**Accuracy**

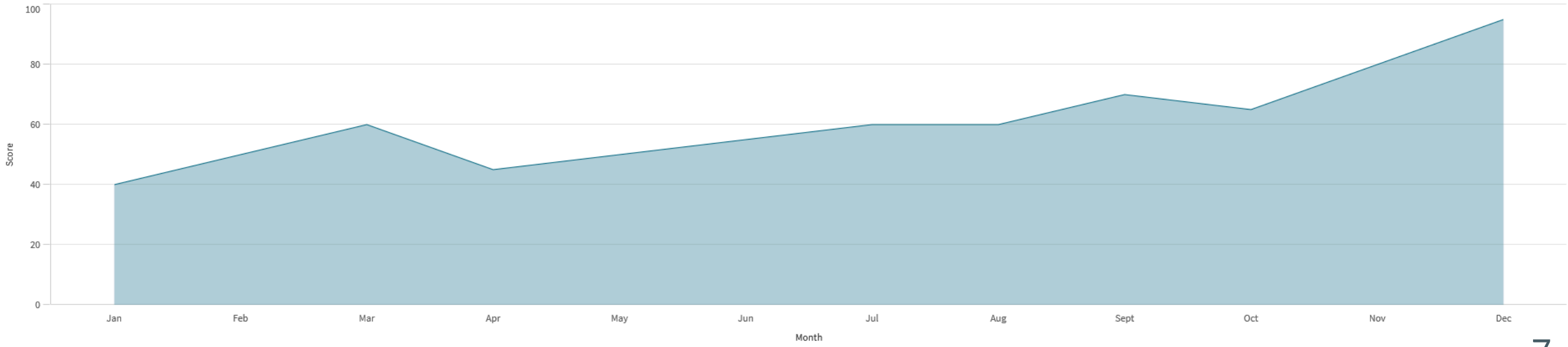



**Consistency**


**Uniqueness**



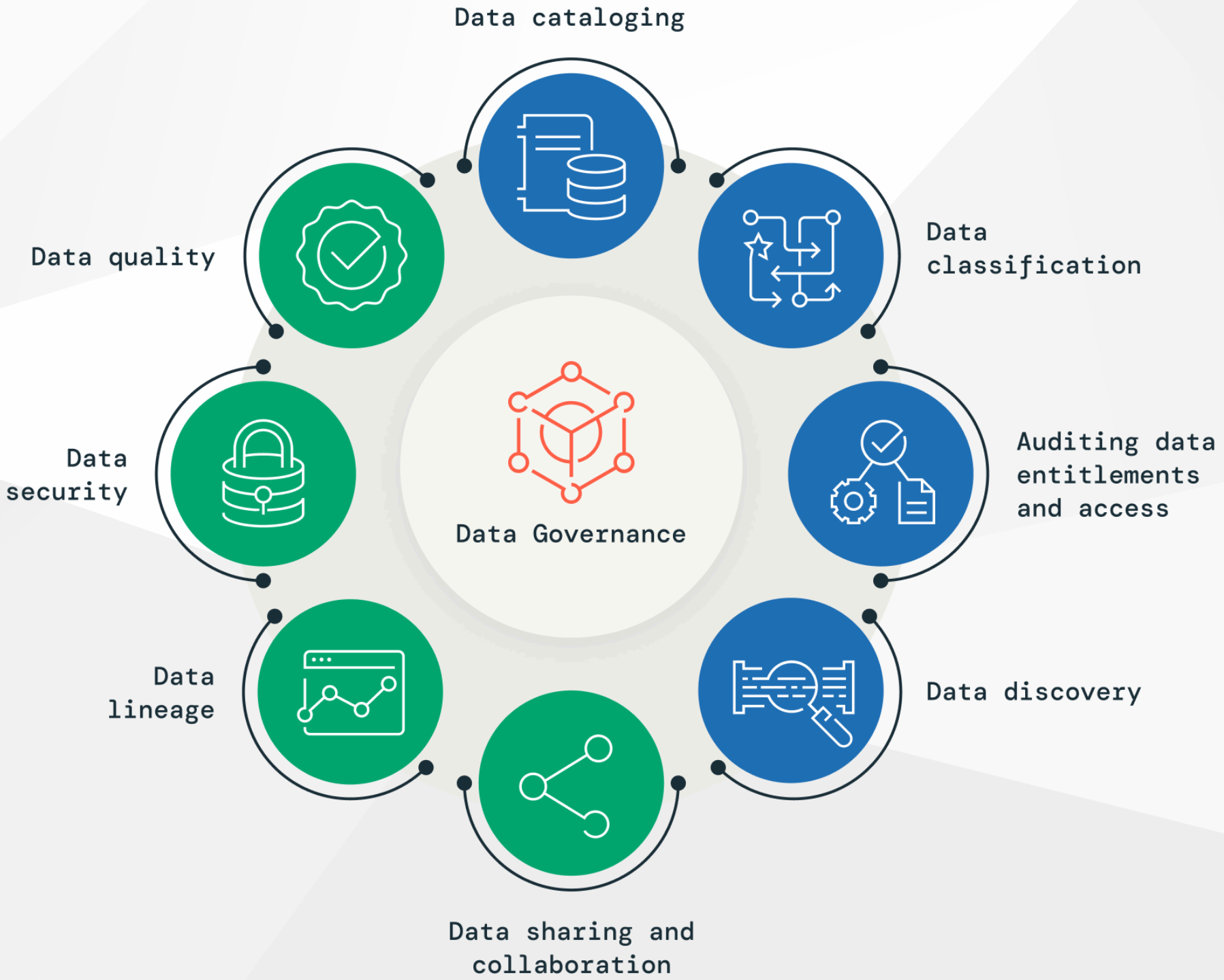

**Overall DQ score - last 12 months**



# Roles básicos en DQA

- Productor de datos
- Consumidor de datos
- Grupo de gobernanza de datos

Son roles intercambiables y es importante establecer cual sombrero traemos puesto en cada momento.


Data cataloging
Data quality
Data classification
Data security
Auditing data entitlements and access
Data Governance
Data lineage
Data discovery
Data sharing and collaboration

# Conclusión

- La calidad de los datos es esencial para la toma de decisiones informadas y la obtención de resultados confiables.
- Al comprender las dimensiones de la calidad de datos y estableciendo reglas específicas, podemos garantizar que nuestra información sea precisa, completa y confiable.