La Ciencia de Datos en la Institución

Métricas y metodologías ágiles para la gestión de datos

Clase 5

MAESTRÍA EN
SISTEMAS DE INFORMACIÓN
Mención Data Science

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

Métricas y metodologías ágiles para la gestión de datos

En esta clase, se aborda cómo las métricas y las metodologías ágiles, específicamente Scrum, juegan un rol fundamental en la gestión de datos en ciencia de datos. Las métricas de apoyo a la gobernanza de datos permiten evaluar y mejorar la calidad, seguridad, disponibilidad e integridad de los datos, factores clave para asegurar su valor y utilidad en la toma de decisiones. Asimismo, la metodología Scrum brinda una estructura ágil que facilita la organización de los equipos de datos y la adaptación continua a los cambios en los requisitos del proyecto y del negocio.

La sinergia entre métricas de gobernanza de datos y metodologías ágiles permite a las organizaciones optimizar sus prácticas de gestión de datos y asegurar que estos se alineen con sus metas estratégicas. Este capítulo, por tanto, ofrece una visión integral de cómo combinar métricas y agilidad en proyectos de ciencia de datos para maximizar la efectividad en la gestión de datos y mejorar la toma de decisiones en tiempo real.

Indicar los RDAs que corresponden a esta clase:

RDA 2: Aplicar marcos de referencia y metodologías para la gestión y gobierno de datos en ciencia de datos.

Clase 5: Métricas y metodologías ágiles para la gestión de datos

1.1. Métricas de Apoyo a la Gobernanza de Datos

La gobernanza de datos es una práctica esencial que busca asegurar que los datos dentro de una organización sean precisos, seguros y utilizados de manera eficiente para tomar decisiones informadas. Dentro de este contexto, las métricas juegan un rol fundamental al evaluar la efectividad de las prácticas de gobernanza. Estas métricas permiten a las organizaciones monitorear, medir y mejorar continuamente sus estrategias de gestión de datos (DAMA International, 2017).

La gobernanza de datos se puede entender como un sistema de políticas, normas y procesos que tiene como objetivo garantizar que los datos sean gestionados adecuadamente y que respalden las decisiones empresariales de manera confiable. Para medir la efectividad de la gobernanza de datos, las organizaciones emplean métricas específicas que evalúan la calidad, la disponibilidad y el acceso seguro a los datos (Singh & Singh, 2021). Estas métricas permiten una visibilidad completa de los datos, algo crucial en el entorno de datos masivos y sistemas complejos actuales. Como menciona Marr (2015), en el contexto de Big Data y analítica avanzada, las métricas son imprescindibles para generar valor a partir de los datos y no solo para almacenarlos.

En su obra sobre gobernanza de datos, Singh y Singh (2021) destacan que la implementación de métricas proporciona una base cuantitativa para evaluar el rendimiento de la gobernanza de datos. Esto ayuda a que los responsables de la toma de decisiones comprendan el estado actual de sus datos y las áreas donde se requieren mejoras. El DAMA-DMBOK (Data Management Body of Knowledge), por ejemplo, establece categorías clave para estas métricas, como calidad de los datos, disponibilidad y accesibilidad, que permiten a las organizaciones medir los elementos críticos de sus iniciativas de gobernanza (DAMA International, 2017).

1.1.1. Principales Métricas para la Gobernanza de Datos (Figura 1)

Figura 1 Métricas para la Gobernanza de Datos

1.1.1.1. Calidad de los Datos

La calidad de los datos es una de las métricas más importantes en la gobernanza de datos. Incluye atributos como la precisión, integridad, consistencia, actualidad y relevancia de los

datos. Para que los datos sean útiles en la toma de decisiones, deben ser exactos y estar libres de errores. Como señala DAMA International (2017), la precisión es crucial, ya que errores en los datos pueden llevar a decisiones equivocadas, impactando negativamente en los objetivos organizacionales. En un estudio sobre la calidad de los datos, Nagle, Sammon y Brennan (2020) explican que la integridad de los datos también juega un papel fundamental, ya que garantiza que los datos estén completos y no se pierdan al trasladarse entre sistemas.

La relevancia de la calidad de los datos se observa especialmente en el análisis de Big Data, donde la cantidad de datos disponibles no necesariamente se traduce en información útil. En este contexto, las métricas de calidad ayudan a filtrar y mejorar los datos antes de que sean usados para generar insights (Marr, 2015).

1.1.1.2. Disponibilidad de los Datos

La disponibilidad de los datos se refiere a la accesibilidad de los datos cuando los usuarios los necesitan. Esta métrica es fundamental para evitar interrupciones en el flujo de trabajo y asegurar que los datos estén accesibles en el momento adecuado. Singh y Singh (2021) sostienen que la falta de disponibilidad de los datos puede llevar a demoras en la toma de decisiones, afectando la eficiencia operativa. Las organizaciones deben implementar políticas y tecnologías que garanticen que sus datos estén disponibles continuamente, considerando factores como la redundancia y la recuperación ante desastres.

Para medir la disponibilidad, se utilizan métricas que evalúan el tiempo de inactividad de los sistemas de datos y la velocidad de recuperación después de fallas. En este sentido, la DAMA-DMBOK recomienda tener en cuenta tanto la accesibilidad de los datos como la facilidad de acceso a estos, para asegurar una disponibilidad adecuada (DAMA International, 2017).

1.1.1.3. Integridad de los Datos

La integridad se refiere a la precisión y consistencia de los datos durante todo su ciclo de vida. Esta métrica es esencial para garantizar que los datos no se alteren o corrompan, especialmente durante procesos de transferencia o migración entre sistemas. Según Singh y Singh (2021), la integridad se mantiene mediante controles y políticas que aseguran la precisión de los datos a medida que se mueven entre sistemas o son transformados.

Las organizaciones pueden medir la integridad de los datos mediante herramientas de auditoría y control de calidad que identifican inconsistencias o anomalías en los datos. La DAMA-DMBOK sugiere que la integridad de los datos debe verificarse regularmente para garantizar que los datos permanecen fiables a lo largo de sus transformaciones y transferencias (DAMA International, 2017).

1.1.1.4. Seguridad de los Datos

La seguridad de los datos es una métrica crítica en la gobernanza de datos, dado que protege la confidencialidad y la integridad de la información sensible. Las brechas de seguridad pueden tener consecuencias graves, tanto financieras como reputacionales, para las organizaciones. En la obra de Singh y Singh (2021), se subraya la importancia de implementar políticas de seguridad que incluyan autenticación, autorización y encriptación, para asegurar que solo los usuarios autorizados puedan acceder a la información.

Las métricas de seguridad suelen centrarse en el número de incidentes de seguridad, el tiempo de respuesta a estos incidentes y la efectividad de los sistemas de seguridad implementados. Para las organizaciones, es fundamental contar con mecanismos de monitoreo continuo que permitan detectar y responder a amenazas de manera rápida y efectiva (Schwaber & Sutherland, 2020).

1.1.1.5. Cumplimiento Normativo

El cumplimiento normativo es otra métrica clave en la gobernanza de datos, ya que asegura que las prácticas de manejo de datos cumplan con las normativas y regulaciones vigentes. Esto es particularmente relevante en sectores altamente regulados, como el financiero o el de la salud. Según Klipp y Lubke (2013), el cumplimiento normativo no solo es una obligación legal, sino que también ayuda a proteger la reputación de la empresa.

Las métricas de cumplimiento incluyen auditorías periódicas y el seguimiento de las políticas de privacidad y seguridad de datos. Estas métricas ayudan a las organizaciones a asegurar que sus prácticas de manejo de datos estén alineadas con las regulaciones y a identificar posibles áreas de riesgo. La DAMA-DMBOK sugiere que el cumplimiento normativo debe integrarse en las políticas de gobernanza de datos desde el inicio, para evitar costosas sanciones y proteger la imagen de la empresa (DAMA International, 2017).

Implementación y Evaluación de Métricas de Gobernanza de Datos

Las métricas de gobernanza de datos requieren un sistema de monitoreo constante para evaluar el rendimiento en tiempo real y realizar ajustes oportunos. Según Cohn (2009), la evaluación continua asegura que los datos cumplan con los estándares de calidad de la organización. Además, definir roles y responsabilidades en la supervisión de métricas como calidad, disponibilidad y seguridad es clave para una gestión efectiva (Leffingwell, 2018). Estos elementos no solo garantizan datos precisos y seguros, sino también el cumplimiento de normativas, optimizando la toma de decisiones y maximizando el valor de los datos organizacionales.

Enlace relacionado.

- **7 métricas de calidad de datos para evaluar el estado de sus datos**
  Enlace: https://www.astera.com/es/type/blog/data-quality-metrics/
  En este artículo se discuten las métricas de calidad en Big Data, destacando cómo la precisión, disponibilidad y seguridad de los datos afectan los proyectos de análisis de datos.

1.2. Metodologías Ágiles: Scrum

Actualmente, las metodologías ágiles son esenciales en la gestión de proyectos de datos que requieren flexibilidad y adaptabilidad. Entre ellas, destaca Scrum por su enfoque en colaboración, iteración continua y entrega constante de valor (Cohn, 2009). Scrum permite gestionar proyectos de ciencia de datos mediante roles, eventos y artefactos clave, y aporta beneficios frente a metodologías tradicionales al mejorar la gobernanza y calidad de los datos en entornos complejos.

Scrum, que fue usado inicialmente en desarrollo de software, es eficaz en proyectos de datos al organizar el trabajo en sprints, ciclos iterativos de dos a cuatro semanas, que entregan resultados funcionales (Schwaber & Sutherland, 2020). Este marco se estructura en tres roles principales—Scrum Master, Product Owner y equipo de desarrollo—y emplea eventos y artefactos que optimizan la transparencia y organización del proceso, permitiendo al equipo adaptarse y mejorar continuamente, lo cual es esencial en proyectos de datos (Cohn, 2009).

Enlace relacionado.

- **Scrum y Metodologías Ágiles en la Ciencia de Datos**
  Enlace: https://institutohumai.medium.com/metodolog%C3%ADas-%C3%A1giles-en-machine-learning-3a68a17ac97f
  En este enlace se explica los principios fundamentales de las metodologías ágiles y cómo estas metodologías nos ayudan para el desarrollo de proyectos de Machine Learning.

1.2.1. Roles en Scrum (Figura 2)

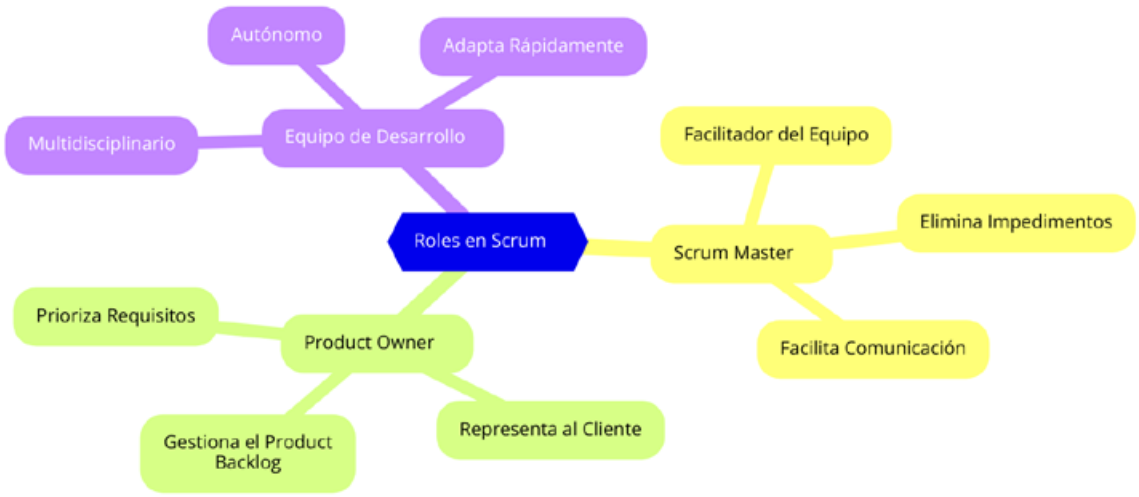

Figura 2 Roles en Scrum

1.2.1.1. Scrum Master

El Scrum Master actúa como facilitador y líder del equipo, asegurándose de que se sigan los principios y prácticas de Scrum. Aunque no es un rol de autoridad, el Scrum Master es responsable de guiar al equipo y eliminar cualquier impedimento que pueda afectar el progreso del proyecto. En el contexto de un proyecto de datos, el Scrum Master juega un papel crucial para mantener la coherencia del equipo y garantizar que los miembros puedan trabajar de manera ágil y colaborativa (Schwaber & Sutherland, 2020).

Como indica Leffingwell (2018), el rol del Scrum Master también implica facilitar la comunicación entre el equipo de datos y otras áreas de la organización, promoviendo la colaboración y la comprensión de los objetivos del proyecto. En un entorno de ciencia de datos, el Scrum Master es clave para alinear las expectativas de los stakeholders y para asegurar que los datos producidos cumplan con los requisitos de calidad y utilidad.

1.2.1.2. Product Owner

El Product Owner es el representante de los intereses del cliente y de los stakeholders. Es el encargado de definir y priorizar los requisitos del proyecto, que se plasman en un documento conocido como *product backlog*. En el contexto de un proyecto de ciencia de datos, el Product Owner debe tener una comprensión clara de las necesidades de los usuarios y de los objetivos del negocio, y debe ser capaz de traducir estas necesidades en requerimientos específicos de datos (Schwaber & Sutherland, 2020).

En su obra, Cohn (2009) subraya que el Product Owner también debe mantener una comunicación constante con el equipo para ajustar las prioridades en función de los avances y las nuevas necesidades que puedan surgir. En proyectos de datos, este rol es especialmente importante, ya que los requisitos pueden cambiar rápidamente en respuesta a nuevos hallazgos o a cambios en el contexto del negocio. El Product Owner debe equilibrar la visión de largo plazo con la flexibilidad para hacer ajustes según sea necesario.

1.2.1.3. Equipo de Desarrollo

El equipo de desarrollo en Scrum está compuesto por profesionales multidisciplinarios que son responsables de realizar el trabajo necesario para cumplir con los objetivos del sprint. En un proyecto de datos, este equipo puede incluir científicos de datos, ingenieros de datos, analistas y otros especialistas que contribuyen a la creación de productos de datos de alta calidad (Schwaber & Sutherland, 2020).

Este equipo es autónomo y se organiza de manera independiente, lo que le permite adaptarse rápidamente a cambios en los requisitos del proyecto. Según Klipp y Lubke (2013), la autonomía del equipo es una de las características que permiten a Scrum responder eficazmente a la complejidad y a la velocidad de los proyectos de datos. Cada miembro del equipo aporta habilidades complementarias, lo cual es esencial para resolver los desafíos técnicos y analíticos que suelen surgir en proyectos de datos.

1.2.2. Eventos en Scrum

1.2.2.1. Sprint

El sprint es el núcleo del marco de Scrum. Cada sprint tiene una duración fija y representa un ciclo de trabajo completo que termina con la entrega de un producto funcional o un incremento del producto final. La duración del sprint permite al equipo enfocarse en objetivos específicos y evitar cambios que puedan desviar su atención, promoviendo la consistencia y la entrega continua de valor (Schwaber & Sutherland, 2020).

En el contexto de la ciencia de datos, los sprints pueden utilizarse para completar tareas específicas, como la limpieza de datos, el desarrollo de modelos predictivos o la generación de informes. La duración del sprint permite que el equipo ajuste sus esfuerzos en función de los resultados obtenidos en cada iteración, facilitando una evolución constante del proyecto y una adaptación rápida a los cambios en los datos o en los requisitos (Cohn, 2009).

1.2.2.2. Reunión de Planificación del Sprint

Esta reunión marca el inicio de cada sprint y en ella se definen los objetivos específicos que el equipo buscará alcanzar durante el ciclo. En este evento, el Product Owner presenta las prioridades del backlog y el equipo de desarrollo decide qué elementos pueden abordarse en el sprint actual (Schwaber & Sutherland, 2020).

La planificación es crítica en proyectos de datos debido a la necesidad de definir claramente los resultados y los objetivos de cada sprint. En palabras de Leffingwell (2018), una planificación adecuada garantiza que el equipo tenga un entendimiento claro de las tareas que deben realizarse y reduce las posibilidades de desvíos o malentendidos.

1.2.2.3. Revisión del Sprint

Al finalizar cada sprint, el equipo realiza una reunión de revisión en la que presenta los re-

sultados obtenidos y recibe retroalimentación de los stakeholders y del Product Owner. Esta reunión permite ajustar el proyecto de manera continua y asegura que el trabajo realizado cumpla con las expectativas y requisitos del negocio (Schwaber & Sutherland, 2020).

En proyectos de ciencia de datos, la revisión del sprint es especialmente útil para ajustar modelos, corregir errores y mejorar los resultados en función de la retroalimentación recibida. Según Cohn (2009), este evento es una oportunidad para identificar mejoras que pueden aplicarse en los siguientes sprints, fomentando una evolución constante del producto de datos.

1.2.2.4. Retrospectiva del Sprint

La retrospectiva es un evento interno que permite al equipo reflexionar sobre el proceso de trabajo y buscar formas de mejorar. Durante esta reunión, el equipo discute lo que funcionó bien, lo que podría mejorar y cualquier cambio que necesite implementar en el siguiente sprint (Schwaber & Sutherland, 2020).

En el contexto de la gestión de datos, la retrospectiva permite que el equipo aborde problemas específicos relacionados con la calidad y la gestión de los datos. Klipp y Lubke (2013) señalan que, en proyectos de datos, esta reflexión es fundamental para ajustar procesos, mejorar la precisión y calidad de los modelos analíticos y optimizar la colaboración entre los miembros del equipo.

1.2.3. Artefactos en Scrum

1.2.3.1. Product Backlog

Es una lista priorizada de los requisitos necesarios para completar el proyecto, y se ajusta constantemente según las necesidades del negocio y las prioridades del proyecto (Schwaber & Sutherland, 2020). En proyectos de datos, el backlog incluye tareas específicas, como la adquisición, limpieza, modelado y visualización de datos. Esta adaptabilidad permite al *Product Owner* y al equipo de datos responder rápidamente a nuevos requisitos o descubrimientos en el análisis (Cohn, 2009).

1.2.3.2. Sprint Backlog

Es una selección del *product backlog* que contiene los elementos específicos para trabajar en un sprint. Ayuda a mantener el enfoque del equipo, proporcionando una guía clara de las tareas necesarias para cumplir los objetivos de ese ciclo (Schwaber & Sutherland, 2020). En proyectos de datos, permite abordar tareas específicas del ciclo de vida de los datos, como la limpieza, el desarrollo de modelos y validación, optimizando el flujo de trabajo (Leffingwell, 2018).

1.2.3.3. Incremento de Producto

Es el resultado acumulativo de los sprints realizados. Cada incremento debe ser funcional y listo para entrega al usuario final, cumpliendo con los estándares de calidad definidos (Schwaber & Sutherland, 2020). En proyectos de ciencia de datos, esto puede incluir un modelo predictivo, un análisis o una visualización interactiva que aporte valor a los stakeholders y permita decisiones informadas (Cohn, 2009).

1.2.4. Beneficios de Utilizar Scrum en la Gestión de Datos

Scrum ofrece importantes beneficios para proyectos de datos, mejorando adaptabilidad, rapidez y

colaboración, lo que lo convierte en una metodología ideal en ciencia de datos, donde los requisitos cambian frecuentemente y la entrega rápida de resultados es crítica. La jerarquía de beneficios se puede ver en la Figura 3.

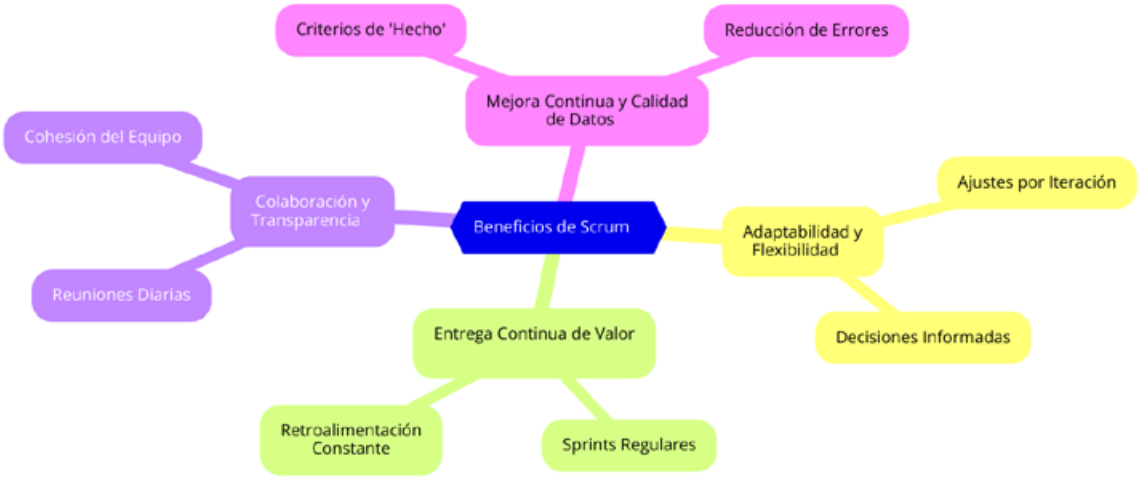

Figura 3 **Beneficios de Utilizar Scrum en la Gestión de Datos**

1.2.4.1. Adaptabilidad y Flexibilidad

Scrum permite ajustes en cada iteración, adaptándose a cambios en el proyecto o en el contexto de los datos, lo cual es clave para decisiones informadas Leffingwell (2018).

1.2.4.2. Enfoque en la Entrega Continua de Valor

Con sprints regulares, los stakeholders reciben resultados tangibles y pueden realinear el proyecto según retroalimentación constante (Schwaber & Sutherland, 2020).

1.2.4.3. Colaboración y Transparencia

Las reuniones diarias y retrospectivas facilitan la cohesión y superación de desafíos técnicos (Cohn, 2009).

1.2.4.4. Enfoque en la Mejora Continua y Calidad de Datos

La revisión en cada sprint y los criterios de "hecho" aseguran estándares altos desde el inicio, reduciendo errores y fortaleciendo la precisión de los resultados Singh y Singh (2021).

1.2.5. Retos de Implementar Scrum en Proyectos de Datos

Implementar Scrum en proyectos de datos presenta desafíos específicos debido a la complejidad y variabilidad inherentes a los datos y a la necesidad de alinear equipos multidisciplinarios.

1.2.5.1. Variabilidad en los Requisitos de Datos

Los requisitos pueden cambiar drásticamente debido a nuevos hallazgos o prioridades del negocio, lo cual dificulta cumplir los objetivos del sprint sin una estrategia sólida para manejar estos cambios (Leffingwell, 2018).

1.2.5.2. Complejidad Técnica y Análisis de Datos

Tareas como procesamiento masivo de datos o desarrollo de modelos predictivos requieren tiempo y habilidades avanzadas; la iteración rápida de Scrum puede ser insuficiente para procesos que necesitan precisión (Cohn, 2009).

1.2.5.3. Coordinación de Equipos Multidisciplinarios

Equipos con expertos en ciencia de datos, ingeniería y negocio requieren una coordinación eficaz, ya que cada disciplina tiene expectativas y enfoques distintos (Schwaber & Sutherland, 2020).

1.2.5.4. Manejo de la Calidad y el "Hecho" en el Contexto de Datos

La naturaleza exploratoria dificulta definir criterios de "hecho" claros, lo cual puede afectar la calidad y entrega del producto en cada sprint (Singh & Singh, 2021).

Scrum es una metodología ágil eficaz en la gestión de proyectos de ciencia de datos, ofreciendo flexibilidad y un enfoque en la entrega continua de valor. A través de roles, eventos y artefactos, facilita la colaboración y transparencia, esenciales en la gestión de datos. Sus ciclos iterativos permiten que los equipos se adapten a cambios y mejoren continuamente los resultados, lo cual es clave en entornos de datos dinámicos. Sin embargo, su implementación en ciencia de datos plantea retos, como la variabilidad en los requisitos y la complejidad técnica. Para superarlos, los equipos deben definir estrategias y criterios claros de "hecho". La combinación de métricas de gobernanza y Scrum permite a las organizaciones mejorar la calidad de sus datos y adaptarse rápidamente a cambios, fortaleciendo así su ventaja competitiva en un mercado en evolución.

Referencias citadas en la Clase 5.

Cohn, M. (2009). *Succeeding with Agile: Software Development Using Scrum*. Addison-Wesley Professional.

DAMA International. (2017). *DAMA-DMBOK: Data Management Body of Knowledge (2nd Edition)*. Technics Publications.

Klipp, P., & Lubke, T. (2013). *The Agile Data Warehouse Design: Collaborative Dimensional Modeling, from Whiteboard to Star Schema*. Addison-Wesley Professional.

Leffingwell, D. (2018). *SAFe 4.5 Reference Guide: Scaled Agile Framework for Lean Enterprises*. Addison-Wesley Professional.

Schwaber, K., & Sutherland, J. (2020). *The Scrum Guide: The Definitive Guide to Scrum: The Rules of the Game*. Scrum.org.

Singh, A., & Singh, A. P. (2021). *Data Governance: The Definitive Guide: People, Processes, and Tools to Operationalize Data Trustworthiness*. O'Reilly Media.

Definición de los términos citados en la Clase 5.

1. **Calidad de los Datos**: Métrica en la gobernanza de datos que mide atributos como la precisión, integridad y consistencia de los datos, esenciales para asegurar su utilidad en la toma de decisiones (DAMA International, 2017).

2. **Cumplimiento Normativo**: Métrica de gobernanza que evalúa la alineación de las prácticas de manejo de datos con las normativas y regulaciones vigentes, minimizando riesgos legales y reputacionales (Singh & Singh, 2021).

PUCE Virtual

La excelencia no se improvisa

síguenos

MDASC - 12 - 241