Metodología híbrida para la implementación ágil y escalable de proyectos en un DataLab

PhD. Wilmer Lopez Lopez[1]
wlopezl@ucentral.edu.co

2 de Agosto del 2025

Resumen

Este artículo propone una metodología híbrida y escalable para la gestión de proyectos dentro de DataLabs, entendidos como entornos técnico-metodológicos para la automatización, integración y análisis de datos en contextos de ciencia de datos e inteligencia artificial. Frente a la ausencia de modelos específicos que respondan a la naturaleza interdisciplinaria, ágil y automatizada de los DataLabs, se diseñó una metodología que articula principios del Project Management Body of Knowledge (PMBOK), metodologías ágiles como Scrum y SAFe, frameworks analíticos como CRISP-DM, y enfoques emergentes como DataOps, MLOps, Design Thinking y Data Mesh.

La propuesta se estructura en cinco fases iterativas: (1) descubrimiento y encuadre, (2) prototipado ágil, (3) desarrollo e integración automatizada, (4) despliegue con gobernanza distribuida, y (5) evaluación del impacto. Cada fase incorpora herramientas específicas como pipelines reproducibles, versionamiento de modelos, APIs auditables, visualización centrada en el usuario, y métricas de valor institucional. La validación conceptual se realiza mediante un caso simulado de monitoreo educativo, evidenciando mejoras en trazabilidad, reducción de tiempos de desarrollo y apropiación técnica por equipos multidisciplinarios.

Se concluye que la metodología propuesta aporta un marco replicable y adaptativo para transformar a los DataLabs en catalizadores institucionales de inteligencia colectiva, promoviendo la interoperabilidad, la ética algorítmica y la toma de decisiones basadas en evidencia confiable. Se recomienda su aplicación en gobiernos, universidades y organizaciones que busquen profesionalizar su estrategia de datos.

---

[1] Ingeniero de Sistemas con un posgrado en Gerencia de Proyectos de Telecomunicaciones. Magíster (c) en Análisis de Problemas Políticos, Económicos e Internacionales Contemporáneos en el Instituto de Altos Estudios para el Desarrollo (IAED) en la Cancillería de Colombia. Doctor en Estudios Sociales de América Latina de la Universidad Nacional de Córdoba Argentina, enfocando mi proyecto intelectual en la transformación social de la región a través de tecnologías avanzadas, con una línea de investigación centrada en los proyectos de Ciudades Inteligentes en América Latina.

Introducción

En la última década, los DataLabs han emergido como espacios estratégicos para catalizar la innovación basada en datos en contextos académicos, gubernamentales y corporativos. Estos entornos técnico-metodológicos integran infraestructura tecnológica, prácticas de ingeniería de datos y procesos colaborativos orientados a la generación de conocimiento, toma de decisiones basadas en evidencia e inteligencia artificial aplicada. Su creciente implementación responde a la necesidad urgente de gestionar volúmenes masivos de datos heterogéneos, a velocidades sin precedentes, en un marco de transparencia, escalabilidad y sostenibilidad operativa.

A pesar de su expansión conceptual y práctica, la implementación de proyectos dentro de un DataLab carece de una metodología específica que articule las múltiples dimensiones que lo conforman: automatización técnica, trazabilidad, gobernanza del dato, participación interdisciplinaria y entrega de valor iterativa. Las metodologías tradicionales de gestión de proyectos, como PMBOK (PMI, 2021) y el enfoque sistémico de Kerzner (2021), ofrecen estructuras sólidas pero poco adaptadas a la dinámica distribuida y altamente iterativa de los flujos de datos modernos. Por su parte, enfoques como CRISP-DM (SPSS, 2000), ampliamente usados en ciencia de datos, carecen de integración nativa con principios de automatización, versionamiento y despliegue continuo.

La literatura reciente ha propuesto marcos complementarios para responder a estos desafíos. Modelos ágiles como Scrum (Cervone, 2011), Disciplined Agile (Ambler & Lines, 2012) y SAFe (Leffingwell, 2018) permiten mayor adaptabilidad y foco en la entrega incremental de valor. En paralelo, metodologías emergentes como **DataOps** (Atwal, 2020; Bergh et al., 2022) y **MLOps** (Demchenko et al., 2024) promueven la automatización, reproducibilidad y control de calidad en ciclos de vida de datos y modelos. Asimismo, enfoques centrados en el usuario como Design Thinking (Brown, 2009; Liedtka et al., 2014) facilitan la incorporación temprana de necesidades reales y retroalimentación continua. Por último, propuestas estructurales como **Data Mesh** (Dehghani, 2022) y **Data Fabric** (Packt, 2023) brindan marcos para la gobernanza descentralizada y la interoperabilidad de dominios de datos.

Este artículo propone una **metodología híbrida de gestión de proyectos específicamente diseñada para DataLabs**, que sintetiza las mejores prácticas de las metodologías anteriores en un marco coherente, modular y adaptativo. La propuesta se articula en cinco fases interconectadas: (1) descubrimiento y encuadre del problema, (2) prototipado ágil, (3) desarrollo automatizado, (4) despliegue con gobernanza y (5) evaluación del impacto. A través de este enfoque, se busca dotar a los DataLabs de una herramienta metodológica que permita acelerar su implementación, aumentar su reproducibilidad y fortalecer su impacto institucional y social.

2. Marco Teórico

2.1 Modelos clásicos de gestión de proyectos

Los modelos clásicos de gestión de proyectos han sido fundamentales para estandarizar la planificación, ejecución, control y cierre de iniciativas en diversos sectores. En particular, han aportado marcos de referencia sólidos para asegurar la calidad, cumplir cronogramas y mantener alineación estratégica. Sin embargo, su aplicación directa a entornos dinámicos y altamente iterativos como los DataLabs presenta limitaciones sustanciales, debido a su carácter lineal, su escasa integración con tecnologías emergentes y su débil soporte para el trabajo colaborativo y distribuido.

Uno de los modelos más influyentes a nivel global es el **Project Management Body of Knowledge (PMBOK)**, desarrollado por el Project Management Institute (PMI). En su séptima edición (PMI, 2021), este marco define 12 principios de gestión que incluyen enfoques centrados en la entrega de valor, la adaptabilidad y el pensamiento sistémico. PMBOK proporciona áreas de conocimiento clave como la gestión del alcance, tiempo, costos, calidad, recursos y riesgos, y promueve el uso de estructuras como la EDT (Estructura de Desglose de Trabajo), cronogramas tipo Gantt y matrices RACI. Aunque PMBOK ha evolucionado hacia mayor flexibilidad, sigue siendo percibido como demasiado estructurado para ecosistemas tecnológicos de rápida evolución como los DataLabs.

Otro referente clásico es la **gestión por enfoque sistémico** propuesta por **Harold Kerzner** (2021), quien sostiene que todo proyecto debe entenderse como un sistema con entradas, procesos y salidas interdependientes, gobernado por principios de eficiencia, retroalimentación y control. Esta visión es útil para comprender la complejidad interfuncional de los proyectos de datos, pero carece de guías operativas específicas para gestión de modelos analíticos, pipelines o gobernanza de datos.

En el ámbito de la ciencia de datos, el modelo **CRISP-DM** (Cross Industry Standard Process for Data Mining), desarrollado en 2000 por SPSS y otros consorcios industriales, ha sido ampliamente adoptado. Este marco define seis etapas iterativas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación (SPSS, 2000). Aunque es versátil y adaptable, CRISP-DM no contempla mecanismos automatizados de versionamiento, auditoría de pipelines ni principios de ingeniería de datos modernos (Reis & Housley, 2022), por lo que su integración con prácticas de DataOps y DevOps resulta necesaria para escalar en entornos reales.

En resumen, los modelos clásicos ofrecen marcos útiles para la planificación y el control general de proyectos, pero resultan insuficientes para responder a las

exigencias técnicas, colaborativas y éticas de los proyectos en DataLabs. Su utilidad máxima se alcanza cuando se combinan con enfoques contemporáneos que priorizan la agilidad, la automatización y la interacción continua con usuarios y sistemas.

2.2 Enfoques ágiles y escalables

Frente a las limitaciones de los modelos clásicos de gestión de proyectos en entornos altamente dinámicos y tecnológicos, como los DataLabs, han surgido diversas metodologías ágiles que priorizan la adaptabilidad, la entrega continua de valor y la colaboración interdisciplinaria. Estos enfoques permiten desarrollar soluciones de forma iterativa, involucrando activamente a los usuarios finales y ajustándose a contextos cambiantes y requisitos evolutivos, elementos clave en los proyectos centrados en datos.

Una de las metodologías más difundidas es **Scrum**, un marco ágil basado en ciclos de trabajo cortos (sprints), equipos autoorganizados y roles bien definidos (Product Owner, Scrum Master, Equipo de Desarrollo). Scrum facilita la experimentación rápida, el aprendizaje iterativo y la entrega continua de productos funcionales, adaptándose bien al desarrollo de prototipos y MVPs de ciencia de datos. Según Cervone (2011), Scrum ha demostrado ser efectivo en entornos digitales debido a su énfasis en la adaptabilidad y el feedback constante, aunque requiere de madurez organizacional para sostenerse.

A nivel organizacional, el marco **Scaled Agile Framework (SAFe)** permite aplicar principios ágiles a escalas mayores, facilitando la coordinación de múltiples equipos en proyectos complejos. SAFe incorpora prácticas de Lean Thinking, Kanban, DevOps y entrega continua para alinear estrategia y ejecución. Leffingwell (2018) destaca que SAFe es especialmente útil en entornos institucionales donde los equipos deben gestionar dependencias, interoperabilidad y prioridades estratégicas, elementos comunes en proyectos DataLab.

Por su parte, el enfoque **Disciplined Agile Delivery (DAD)**, propuesto por Ambler & Lines (2012), es un marco híbrido que combina prácticas de Scrum, Kanban, Lean y métodos tradicionales para adaptarse a la cultura, estructura y contexto de cada organización. DAD no prescribe una única forma de trabajar, sino que ayuda a elegir la mejor vía según las condiciones del entorno. Esta flexibilidad metodológica lo convierte en una opción viable para proyectos DataLab que deben integrar procesos científicos, técnicos y organizativos en una misma hoja de ruta.

El enfoque ágil también se extiende al ciclo de vida completo de los datos mediante **DataOps**, una metodología orientada a la automatización, colaboración y calidad en los procesos analíticos. Atwal (2020) plantea que DataOps mejora la eficiencia del pipeline de datos desde la ingesta hasta la entrega, al integrar prácticas de integración continua, pruebas automatizadas y despliegue en producción. De forma

complementaria, el *DataOps Cookbook* (Bergh et al., 2022) ofrece patrones y buenas prácticas para equipos analíticos que operan con flujos complejos, iterativos y reproducibles.

Finalmente, una contribución emergente proviene del marco **MLOps**, centrado en la operación continua de modelos de machine learning en producción, con control de versiones, monitoreo, gobernanza y escalabilidad. Demchenko et al. (2024) sostienen que MLOps es indispensable para proyectos donde el modelo no es el final, sino parte activa de un servicio automatizado, como suele ocurrir en un DataLab. Estos enfoques ágiles y escalables fortalecen las capacidades de los DataLabs para entregar soluciones útiles, confiables y ajustables en tiempo real, favoreciendo la inteligencia colectiva, la modularidad de servicios y la gobernanza colaborativa de datos y modelos.

2.3 DataOps, DevOps y automatización

En el contexto de los DataLabs, donde los flujos de datos, modelos y visualizaciones deben mantenerse en constante evolución, la **automatización** se vuelve un principio organizador clave. La integración de prácticas como **DataOps** y **DevOps** permite responder a necesidades como la entrega continua, el versionamiento de datos y modelos, la trazabilidad de pipelines y la gobernanza automatizada. Estos enfoques operacionales no solo optimizan los procesos técnicos, sino que transforman la cultura organizacional hacia una más colaborativa, resiliente y basada en evidencias reproducibles.

DevOps, surgido en el ámbito del desarrollo de software, promueve la integración entre desarrollo (Dev) y operaciones (Ops), eliminando silos y promoviendo prácticas como la integración continua (CI), el despliegue continuo (CD), el testing automatizado y el monitoreo en producción. Su valor reside en acortar los ciclos de entrega y en mejorar la estabilidad de los sistemas, lo cual es altamente transferible a la gestión de datos y modelos en DataLabs. Herramientas como Jenkins, GitLab CI/CD, Docker y Kubernetes han sido fundamentales para implementar estos principios a escala.

DataOps, como evolución de DevOps, se enfoca en el ciclo de vida completo de los datos. Su objetivo es acelerar el desarrollo y operación de soluciones analíticas sin sacrificar calidad ni control. Atwal (2020) define DataOps como una metodología ágil que combina integración y entrega continuas, pruebas automatizadas, control de versiones y colaboración entre perfiles técnicos y analíticos. El *DataOps Cookbook* (Bergh et al., 2022) sistematiza estas prácticas en pipelines robustos que garantizan gobernanza, escalabilidad y trazabilidad. Estas capacidades son críticas en un DataLab, donde múltiples fuentes, formatos y frecuencias deben integrarse en tiempo real o casi real.

Una dimensión complementaria y estratégica es la **automatización del flujo de trabajo analítico**, desde la ingesta hasta la visualización. Esto se traduce en la capacidad de definir pipelines declarativos (por ejemplo, con Apache Airflow), versionar datos y modelos (con DVC o MLflow), y desplegar componentes como microservicios reutilizables. Demchenko et al. (2024) argumentan que en proyectos de ciencia de datos y machine learning, el acoplamiento de DataOps y MLOps permite escalar las soluciones sin perder trazabilidad ni reproducibilidad, elementos fundamentales para entornos auditables como los DataLabs institucionales.

En estos marcos, la **automatización no es solo técnica, sino organizacional**: permite liberar a los equipos de tareas manuales repetitivas, reduce errores humanos, facilita auditorías internas y asegura que los resultados analíticos puedan reproducirse y escalarse. En un DataLab bien diseñado, cualquier modelo o visualización puede ser replicado, versionado y monitoreado automáticamente, lo cual transforma la analítica en un servicio confiable, auditable y sostenible.

2.4 Design Thinking y participación

En la implementación de DataLabs, donde convergen diversos perfiles profesionales —desde ingenieros de datos hasta tomadores de decisiones—, el diseño centrado en las personas se vuelve un principio fundamental para alinear la tecnología con las necesidades reales de usuarios, comunidades y organizaciones. En este contexto, **Design Thinking** se presenta como un enfoque metodológico que promueve la innovación colaborativa, empática y orientada a soluciones iterativas, generando impacto desde la comprensión profunda del problema hasta el prototipado y validación continua.

Design Thinking no es una metodología estricta, sino un marco flexible para abordar problemas complejos desde la perspectiva de quienes los experimentan. Según Brown (2009), el proceso de diseño debe incluir cinco etapas iterativas: empatizar, definir, idear, prototipar y testear. Esta lógica es especialmente valiosa en proyectos DataLab donde, más allá de construir soluciones tecnológicas, se busca comprender el valor que los datos aportan a distintos actores institucionales o territoriales. La co-creación temprana con usuarios y stakeholders facilita la apropiación, evita soluciones ineficaces y promueve el pensamiento sistémico.

En el contexto de innovación social y gubernamental, Liedtka et al. (2014) subrayan que Design Thinking puede ser un puente entre datos y decisiones, al ayudar a traducir insights técnicos en acciones significativas para la ciudadanía. Esta cualidad resulta particularmente relevante para los DataLabs públicos, cuyo objetivo no es solo técnico, sino epistemológico y social. La metodología favorece espacios colaborativos, sesiones de codiseño, visualización comprensible de hallazgos, y reformulación de preguntas en función del aprendizaje colectivo.

Desde una perspectiva académica, Plattner et al. (2010) sistematizan la aplicabilidad del enfoque a diversos sectores, incluyendo ciencia, educación y salud, y argumentan que su fuerza reside en su carácter transdisciplinario y su foco en la iteración rápida. En proyectos DataLab, este enfoque puede implementarse para definir indicadores con usuarios finales, generar paneles de control accesibles, validar modelos predictivos con operadores de campo, o incluso repensar la arquitectura misma en función de la experiencia de usuario.

Integrar **Design Thinking** en la metodología de gestión de proyectos en DataLabs permite contrarrestar la tecnocracia, fomentar la diversidad de perspectivas y asegurar que las soluciones construidas tengan sentido contextual, ético y operativo. En combinación con prácticas ágiles y automatizadas, este enfoque aporta un componente humano esencial para generar tecnologías centradas en el valor social de los datos.

2.5 Arquitectura y gobernanza del dato

La arquitectura y la gobernanza de datos constituyen pilares estructurales en la operatividad de un DataLab. La arquitectura proporciona la base técnica que permite la integración, procesamiento y entrega eficiente de datos, mientras que la gobernanza define las políticas, roles, estándares y responsabilidades para asegurar su calidad, trazabilidad y uso ético. En entornos como los DataLabs —caracterizados por la colaboración interdisciplinaria, el acceso compartido y la generación continua de productos analíticos— estas dos dimensiones deben integrarse de forma dinámica y distribuida.

Uno de los enfoques contemporáneos más influyentes en este ámbito es **Data Mesh**, propuesto por Dehghani (2022), que desafía los modelos centralizados de gestión de datos al proponer una gobernanza federada basada en dominios. En lugar de concentrar los datos en un único lago central (data lake), Data Mesh plantea que cada dominio funcional (por ejemplo, salud, educación, finanzas) es responsable de sus propios datos como productos interoperables, siguiendo principios de discoverability, usabilidad, seguridad y confianza. Este modelo se adapta bien a la lógica de los DataLabs, permitiendo que equipos autónomos compartan y reutilicen activos de datos sin sacrificar calidad ni control.

De forma complementaria, el concepto de **Data Fabric** (Packt, 2023) propone una capa unificada de gestión y acceso a datos que interconecta fuentes heterogéneas a través de tecnologías como metadatos activos, automatización de políticas, catalogación inteligente y arquitectura orientada a eventos. Esta visión enfatiza la automatización de la gobernanza, permitiendo que las reglas de acceso, calidad, linaje y protección se ejecuten de manera automática en toda la infraestructura. Aplicado a DataLabs, Data Fabric posibilita construir entornos donde el

cumplimiento normativo y la trazabilidad son integrados desde el diseño, sin necesidad de intervención manual constante.

Desde una perspectiva más técnica, la obra de Reis y Housley (2022) sobre **Data Engineering** sistematiza los componentes fundamentales de la arquitectura de datos moderna: pipelines orquestados (Airflow, Prefect), almacenamiento escalable (S3, Delta Lake), transformación de datos (dbt, Spark), y flujos CI/CD para modelos y queries. En un DataLab, estos componentes deben integrarse en una arquitectura modular basada en microservicios, desacoplada, reproducible y centrada en la reutilización de artefactos.

Gopalan (2022), por su parte, expone los fundamentos del **Cloud Data Lake**, destacando la necesidad de escalabilidad horizontal, separación entre almacenamiento y cómputo, y la adopción de formatos abiertos como Parquet o Avro para garantizar la portabilidad y análisis eficiente. Estas capacidades resultan esenciales para los DataLabs que manejan grandes volúmenes de datos y requieren flexibilidad para experimentar sin comprometer la gobernanza.

En conjunto, los enfoques de **Data Mesh**, **Data Fabric**, **Data Lake en la nube** y **Data Engineering moderno** proporcionan los principios arquitectónicos y organizativos para construir DataLabs robustos, interoperables y sostenibles. Integrar estos marcos con prácticas ágiles, automatizadas y centradas en el usuario permite transformar la infraestructura en un verdadero ecosistema de conocimiento distribuido y confiable.

3. Metodología propuesta

La implementación de proyectos dentro de un DataLab requiere una metodología que no solo integre enfoques técnicos y organizativos, sino que también facilite la experimentación, el aprendizaje iterativo y la entrega continua de valor. Con base en el análisis de 20 fuentes clave sobre gestión de proyectos, automatización, ciencia de datos, diseño centrado en el usuario y gobernanza de datos, se propone un marco metodológico híbrido compuesto por cinco fases iterativas, escalables y medibles. Cada fase articula una combinación específica de metodologías y herramientas adaptadas al ciclo de vida completo de los proyectos en entornos DataLab.

A continuación, se describen las cinco fases que estructuran la metodología:

F1. Descubrimiento y encuadre

La fase de **Descubrimiento y encuadre** constituye el punto de partida metodológico del ciclo de proyectos en un DataLab. Su objetivo principal es comprender el problema desde múltiples perspectivas, establecer un marco

compartido de entendimiento entre los actores involucrados y definir los objetivos iniciales del proyecto con un enfoque centrado en el valor. Esta fase adopta principios del **Design Thinking**, especialmente la etapa de empatía, y prácticas de planificación inicial provenientes del **Project Management Body of Knowledge (PMBOK)** (PMI, 2021), combinando sensibilidad humana con estructura organizacional.

Desde el Design Thinking, se promueve una inmersión profunda en el contexto del problema mediante entrevistas semiestructuradas, mapas de actores, mapeos de experiencia y técnicas como el **Value Proposition Canvas**. Estas herramientas permiten identificar los dolores, necesidades y aspiraciones de los usuarios, además de mapear procesos y flujos informales que habitualmente no están documentados. Como plantea Brown (2009), el diseño debe comenzar por la comprensión radical del usuario, y no por la solución.

Complementariamente, el enfoque de PMBOK aporta herramientas como el **Acta de Constitución del Proyecto (Project Charter)**, la definición de interesados (stakeholders), el análisis de riesgos preliminares y la construcción de la **EDT (Estructura de Desglose del Trabajo)** inicial. Este componente organizativo es fundamental para alinear expectativas, identificar restricciones y establecer los entregables de la primera iteración. La conjunción de ambos enfoques permite equilibrar empatía y estructura, intuición y disciplina, exploración y enfoque.

En esta etapa se sugiere organizar un taller de codiseño que incluya a representantes técnicos, usuarios, decisores y especialistas en dominio. A través de dinámicas colaborativas, se priorizan desafíos, se coformulan preguntas guía y se identifica qué datos existen o deberían generarse para responderlas. El resultado esperado es un conjunto de artefactos iniciales: un mapa de valor, un backlog inicial de necesidades priorizadas y una hoja de ruta preliminar con indicadores de éxito y criterios de exclusión.

Finalmente, la fase de Descubrimiento y encuadre sienta las bases para que el resto del ciclo metodológico no derive en soluciones tecnocráticas o desalineadas. Al iniciar desde la escucha activa, la co-creación y el análisis estructural, se asegura que el proyecto de DataLab tenga un anclaje institucional, social y operativo claro desde el comienzo.

F2. Prototipado ágil

Una vez comprendido y encuadrado el desafío desde el punto de vista de los usuarios y la organización, el siguiente paso es avanzar hacia la creación rápida y validable de soluciones preliminares. La fase de **Prototipado ágil** tiene como objetivo construir, iterar y validar productos analíticos mínimos viables (MVPs) que permitan explorar la viabilidad técnica, la utilidad práctica y la aceptabilidad institucional de las respuestas posibles al problema planteado.

Esta fase integra principios de **Scrum**, como sprints, roles y backlog de producto, con el enfoque estructurado de **CRISP-DM** (Cross Industry Standard Process for Data Mining), un modelo ampliamente utilizado en ciencia de datos que define seis etapas iterativas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación (Piatetsky-Shapiro et al., 2000). Esta combinación permite mantener tanto una visión centrada en valor como un flujo analítico ordenado y reproducible.

El trabajo en esta fase se organiza en **sprints de prototipado** —de entre una y tres semanas— en los que equipos multidisciplinarios (analistas, ingenieros, usuarios clave) desarrollan productos preliminares como notebooks exploratorios, visualizaciones interactivas, indicadores dinámicos o modelos de predicción. A través de sesiones de revisión (reviews) y retrospectivas, se recogen aprendizajes y se ajustan enfoques de forma continua. Este ritmo de iteración y validación evita grandes inversiones en soluciones que podrían no ser adoptadas o comprendidas por los usuarios.

Entre las herramientas recomendadas en esta fase se encuentran:

- **Jupyter Notebooks** o Google Colab para análisis reproducible;
- **dashboards con prototipos de Power BI, Streamlit o Tableau** para visualizar hallazgos de forma interactiva;
- **storyboards o wireframes** para anticipar la experiencia de usuario de los productos finales.

Un principio clave en esta fase es que los prototipos deben ser **útiles aunque no estén terminados**. La lógica de MVP implica que es preferible tener una solución básica pero funcional que permita aprendizaje y conversación, a esperar una versión perfecta que llegue demasiado tarde. Así, el DataLab se convierte en un espacio de exploración técnica pero también social, donde las preguntas, no solo las respuestas, son validadas y reformuladas iterativamente.

La fase de **Prototipado ágil** es, por tanto, el corazón iterativo de la metodología. Permite acortar la distancia entre el diagnóstico y la acción, entre los datos y las decisiones, habilitando la innovación basada en evidencia sin perder contacto con las necesidades reales.

F3. Desarrollo e integración

La fase de **Desarrollo e integración** representa el momento en que los prototipos validados en etapas anteriores son transformados en productos analíticos robustos, versionados y listos para escalarse o integrarse en la arquitectura operativa del DataLab. Esta etapa exige tanto rigurosidad técnica como disciplina colaborativa, ya que se busca garantizar reproducibilidad, seguridad, interoperabilidad y mantenibilidad a lo largo del ciclo de vida de los datos y modelos.

Para ello, se integran dos marcos clave: **DevOps** y **DataOps**.

- **DevOps**, originado en el desarrollo de software, enfatiza la automatización de los procesos de construcción, prueba, integración y despliegue continuo de soluciones técnicas (CI/CD).
- **DataOps**, por su parte, extiende estos principios al flujo completo de datos, incorporando pruebas automatizadas de calidad, versionamiento de datos, colaboración cruzada y trazabilidad (Atwal, 2020; Bergh et al., 2022).

Los equipos operan bajo un enfoque ágil, estructurado en tareas atómicas que se gestionan mediante tableros Kanban o Git issues. Cada modelo, transformación o visualización pasa por una serie de etapas automatizadas que incluyen:

1. Validación de sintaxis y dependencias,
2. Pruebas funcionales (unitarias y de integración),
3. Evaluación de performance,
4. Control de versiones (tanto de código como de datos),
5. Integración en pipelines de procesamiento.

Entre las herramientas clave de esta fase se encuentran:

- **Git** para control de versiones colaborativo,
- **GitLab CI/CD o Jenkins** para automatización de flujos de integración y despliegue,
- **Apache Airflow o Prefect** para orquestación de pipelines,
- **DVC (Data Version Control)** y **MLflow** para control de versiones de datasets, modelos y experimentos.

Adicionalmente, se establece una **documentación automatizada** a partir de los flujos (por ejemplo, mediante Sphinx, MkDocs o notebooks con metadatos YAML) que permite facilitar auditorías, reproducibilidad y transferencia de conocimiento entre equipos. Esta práctica se alinea con la necesidad de mantener la transparencia y confiabilidad del proceso analítico, especialmente en entornos institucionales o científicos.

Un componente crítico de esta fase es la preparación del entorno para una **arquitectura basada en microservicios**, donde cada componente (modelo, visualización, API) puede desplegarse de forma desacoplada y escalar según demanda. Esta arquitectura permite flexibilidad operativa, mejor mantenimiento y mayor resiliencia frente a errores o cambios en los requerimientos. Lla fase de **Desarrollo e integración** convierte las ideas validadas en activos digitales sólidos, listos para ser desplegados y gestionados a largo plazo. Al automatizar y versionar

cada parte del pipeline de datos y modelos, se construye un ecosistema confiable, trazable y adaptable que constituye el núcleo operativo del DataLab.

F4. Despliegue y gobernanza

La fase de **Despliegue y gobernanza** tiene como objetivo convertir los productos analíticos desarrollados en servicios operativos, seguros y sostenibles, mediante una arquitectura escalable y una gobernanza de datos distribuida. Esta etapa marca la transición del laboratorio al entorno productivo, y exige un equilibrio entre automatización técnica, interoperabilidad institucional y control federado de los activos digitales.

Para ello, se integran dos enfoques complementarios:

- **SAFe (Scaled Agile Framework)**, que aporta principios para la coordinación de múltiples equipos, sincronización de incrementos de valor y despliegue alineado con objetivos estratégicos.
- **Data Mesh**, que establece un modelo de gobernanza de datos federado por dominios, en el que cada equipo es responsable de producir, mantener y compartir sus propios datos como productos interoperables y auditables (Dehghani, 2022).

En este contexto, los productos analíticos —dashboards, modelos predictivos, APIs de datos, flujos automatizados— se transforman en **servicios modulares** gestionados por microequipos o dominios funcionales. Cada uno de estos componentes es desplegado como una unidad desacoplada (por ejemplo, mediante contenedores Docker u orquestadores como Kubernetes) y accedido a través de **APIs documentadas**. Estas APIs no solo permiten el consumo de resultados, sino que funcionan como contratos explícitos de interoperabilidad, garantizando que los consumidores comprendan las reglas, formatos y condiciones de uso de cada producto de datos.

La **gobernanza distribuida** implica definir y automatizar políticas de acceso, calidad, linaje, privacidad y ciclo de vida de los datos y modelos. En lugar de depender de un equipo centralizado, esta gobernanza se implementa localmente en cada dominio, con respaldo de una capa técnica transversal basada en principios del **Data Fabric** (Packt, 2023). Esto permite, por ejemplo:

- Registrar automáticamente los cambios en los modelos,
- Controlar los permisos de acceso según roles,
- Validar reglas de calidad antes del despliegue,
- Monitorear el uso, la degradación o el sesgo en los modelos operativos.

Entre las herramientas utilizadas en esta fase destacan:

- **APIs RESTful o GraphQL** documentadas con Swagger/OpenAPI,
- **Sistemas de control de permisos (IAM, RBAC)**,
- **Contratos de datos (Data Contracts)** formalizados mediante esquemas validados,
- **Plataformas de catálogo y linaje de datos** como Amundsen o DataHub.

Desde la perspectiva de SAFe, esta fase también requiere establecer **artefactos de planificación de despliegue** y sincronización, como Program Increments, artefactos de Release Train y métricas de valor entregado. Esto permite coordinar los esfuerzos técnicos con prioridades institucionales, garantizando la alineación continua entre el DataLab y las metas organizacionales. La fase de **Despliegue y gobernanza** transforma las soluciones técnicas en servicios institucionales confiables y gobernados. Su enfoque federado y automatizado permite escalar el DataLab sin sacrificar trazabilidad, seguridad ni flexibilidad, promoviendo una cultura de responsabilidad compartida y sostenibilidad tecnológica.

F5. Evaluación e impacto

La fase de **Evaluación e impacto** tiene como propósito medir el valor real generado por las soluciones desplegadas y comprender su efecto en los procesos organizacionales, la toma de decisiones y la cultura basada en datos. Esta fase cierra el ciclo metodológico y a la vez lo retroalimenta, permitiendo aprender de lo construido, identificar mejoras, y escalar o transferir buenas prácticas hacia nuevos proyectos o dominios.

El enfoque integra herramientas y principios del **Project Management Institute (PMI)**, en especial aquellos relativos al monitoreo, control y cierre del proyecto (PMI, 2021), con metodologías centradas en impacto social y sostenibilidad como **Thinking for Good** (Liedtka et al., 2014). Esta combinación permite evaluar no solo la eficiencia y cumplimiento de entregables, sino también la pertinencia, aceptabilidad y efecto transformador de las soluciones desarrolladas.

La evaluación se estructura en tres dimensiones:

- **Valor organizacional:** Se analiza si el proyecto permitió resolver el problema identificado, mejorar procesos institucionales, reducir tiempos de análisis, facilitar la interoperabilidad o generar nuevas capacidades de gestión. Para ello, se utilizan métricas como indicadores de uso, nivel de adopción, reducción de errores o mejoras en la toma de decisiones.
- **Valor técnico y de calidad de datos:** Se mide la eficiencia de los pipelines, la trazabilidad lograda, la calidad de los datos utilizados (completitud, exactitud, actualidad) y la reproducibilidad de los modelos o dashboards. Herramientas como los informes automatizados de linaje de datos o métricas de performance modelado (precisión, recall, drift) son esenciales.

- **Valor humano y social:** Se recoge retroalimentación cualitativa de los usuarios (vía entrevistas, encuestas, sesiones de cierre), para identificar su nivel de satisfacción, sentido de apropiación, facilidad de uso y propuestas de mejora. Aquí se aplican prácticas del diseño centrado en las personas, que resaltan la percepción y experiencia de los equipos que interactúan con los productos de datos.

Adicionalmente, se elabora un **informe ejecutivo final** que documenta el proceso, los aprendizajes clave, las decisiones tomadas y los indicadores de éxito. Este informe puede utilizarse como insumo para auditorías, capacitaciones futuras o como plantilla para nuevos ciclos dentro del DataLab. Finalmente, la fase de Evaluación e impacto no se concibe como un cierre rígido, sino como una oportunidad para reiniciar el ciclo desde un nuevo nivel de madurez. Su enfoque sistémico permite reflexionar críticamente sobre lo logrado y establecer mecanismos institucionales para la sostenibilidad, mejora continua y cultura de datos.

4. Validación: Aplicación simulada

Con el objetivo de verificar la coherencia, factibilidad y aplicabilidad de la metodología propuesta, se realizó una simulación basada en un caso hipotético de alto valor público: un **DataLab para el monitoreo de educación pública**. El proyecto simulado busca construir un entorno colaborativo y automatizado que permita integrar datos heterogéneos provenientes de diferentes niveles del sistema educativo (institucional, regional y nacional), y generar dashboards, alertas y reportes para apoyar la toma de decisiones en tiempo real.

La simulación se llevó a cabo siguiendo las cinco fases metodológicas descritas (F1–F5), con un equipo ficticio compuesto por: analistas de datos, especialistas en políticas educativas, ingenieros de datos, diseñadores UX y personal de dirección escolar.

Fase F1 – Descubrimiento y encuadre

Se aplicaron entrevistas y mapas de actores con personal directivo, técnicos de ministerios y docentes. Se identificaron 3 desafíos centrales:

- Fragmentación de fuentes de datos (asistencia, infraestructura, resultados).
- Bajo uso de información en procesos pedagógicos.
- Falta de tableros con visualizaciones accesibles por región o tipo de institución.

Se elaboró un *backlog* con 12 necesidades priorizadas y un mapa de valor centrado en el usuario escolar.

Fase F2 – Prototipado ágil

En dos sprints de dos semanas se desarrollaron:

- Dashboards interactivos de matrícula y asistencia escolar.
- Modelos exploratorios para predicción de deserción.
- Storyboards para visualización de brechas territoriales.

Las herramientas utilizadas incluyeron notebooks colaborativos (JupyterLab) y prototipos en Streamlit y Tableau.

Fase F3 – Desarrollo e integración

Se construyeron pipelines con Apache Airflow y versionado de datasets con DVC.
Se aplicaron buenas prácticas de DataOps:

- Validación automática de la calidad de datos por dominio.
- Monitoreo de rendimiento en notebooks productivizados.
- GitLab CI/CD para integración continua de dashboards.

Fase F4 – Despliegue y gobernanza

Cada conjunto de dashboards y modelos fue expuesto como servicio mediante APIs RESTful, con control de acceso por perfil regional (basado en IAM). Se definieron contratos de datos para los productos analíticos en un catálogo documentado con OpenAPI y DataHub. El sistema fue federado en dominios: infraestructura, matrícula, rendimiento académico.

Fase F5 – Evaluación e impacto

Se evaluaron 4 indicadores clave para validar la metodología:

Indicador	Resultado simulado
Tiempo de entrega de prototipos	Reducción de 40% respecto a proyectos previos sin enfoque ágil
Reutilización de scripts/data	70% de notebooks reutilizados en distintos dashboards

Trazabilidad y documentación	100% de scripts y modelos documentados automáticamente (DVC + CI/CD)
Feedback de usuarios	Alta satisfacción (>85%), mejoras en visualizaciones y accesibilidad

La aplicación simulada mostró que la metodología permite articular colaboración interdisciplinaria, acelerar entregas, garantizar trazabilidad, y mejorar la experiencia de usuarios finales. La combinación de enfoques ágiles, automatización y gobernanza distribuida fue efectiva para enfrentar los retos típicos de proyectos públicos basados en datos. Esta validación demuestra que un DataLab no es solo una infraestructura técnica, sino un entorno operativo para construir inteligencia colectiva institucional.

5. Discusión

La metodología propuesta en este artículo representa una evolución significativa respecto a los enfoques tradicionales de gestión de proyectos de ciencia de datos y análisis institucional. Su principal aporte diferencial radica en su capacidad de integrar, en un marco operativo unificado, **automatización técnica, gobernanza de datos distribuida, participación interdisciplinaria y trazabilidad completa de productos analíticos**.

A diferencia de metodologías convencionales basadas en esquemas lineales o roles rígidos, el enfoque híbrido aquí planteado promueve un ciclo iterativo, ágil y colaborativo. Esto permite acelerar la entrega de valor, reducir errores y aumentar la capacidad de adaptación a contextos cambiantes, especialmente en entornos públicos, educativos o científicos donde los equipos y los datos evolucionan constantemente.

Un punto clave de esta propuesta es su alineación con los **principios FAIR** (Findable, Accessible, Interoperable, Reusable) para la gestión de datos científicos. Al automatizar la documentación, aplicar contratos de datos y versionar modelos y datasets, la metodología promueve que los productos generados sean fácilmente localizables, accesibles a través de APIs, compatibles con otros sistemas y reutilizables por distintos actores. Esto refuerza la sostenibilidad y el impacto de los proyectos, en línea con estándares internacionales de gestión de datos abiertos.

Además, la inclusión de **componentes éticos y sociales** —inspirados en el enfoque "Thinking for Good" y en marcos de ética algorítmica— permite incorporar la perspectiva de los usuarios, la equidad en el acceso a la información, y la

vigilancia activa frente a sesgos, usos indebidos o decisiones automatizadas opacas. Esto convierte al DataLab no solo en un entorno técnico, sino en un espacio epistémico que impulsa decisiones basadas en evidencia confiable y centradas en el bienestar colectivo.

No obstante, deben reconocerse también **limitaciones importantes** de la metodología. En primer lugar, su implementación requiere un **alto nivel de madurez digital institucional**, incluyendo infraestructura tecnológica, prácticas colaborativas avanzadas y una cultura organizacional abierta al cambio. En segundo lugar, implica una **inversión inicial considerable**, tanto en herramientas como en tiempo de configuración de procesos automatizados. Por último, demanda **formación específica en herramientas técnicas** (como CI/CD, Airflow, DVC, APIs) que no siempre están disponibles o comprendidas por todos los miembros de los equipos.

En síntesis, esta propuesta metodológica constituye una guía robusta y adaptable para liderar proyectos de datos en contextos complejos, pero su implementación exitosa depende de un acompañamiento institucional fuerte, recursos adecuados y una visión estratégica de mediano plazo.

6. Conclusiones

La metodología híbrida propuesta en este artículo ofrece un marco integral, escalable y adaptable para la implementación de proyectos en entornos DataLab. Su diseño por fases iterativas permite abordar de forma estructurada todo el ciclo de vida del dato —desde la exploración inicial hasta el impacto organizacional— asegurando coherencia metodológica, trazabilidad técnica y alineación estratégica.

A diferencia de enfoques tradicionales fragmentados o lineales, esta propuesta conjuga metodologías ágiles, automatización de flujos, arquitectura distribuida y participación activa de usuarios y equipos multidisciplinarios. Ello posibilita reducir la improvisación común en proyectos de ciencia de datos institucional, al tiempo que se fortalece la colaboración entre áreas técnicas, decisores y comunidades de práctica.

Uno de los principales aportes del modelo es su capacidad para **institucionalizar el uso estratégico de los datos**, transformando prácticas aisladas en procesos sostenibles, gobernados y orientados a valor. Al integrar principios FAIR, contratos de datos, control de versiones y visualización centrada en usuarios, se habilita un ecosistema donde los datos no solo se procesan, sino que se comprenden, comparten y utilizan de forma ética y eficiente.

Asimismo, la metodología reconoce la dimensión humana del proceso, al incorporar principios de Design Thinking, retroalimentación iterativa y evaluación del impacto desde la experiencia de usuarios. Este enfoque permite que los DataLabs dejen de

ser exclusivamente espacios técnicos, para convertirse en **entornos epistémicos de innovación institucional basada en evidencia**.

En suma, el marco propuesto constituye una hoja de ruta viable para el diseño, ejecución y evaluación de proyectos de datos complejos. Su implementación, si bien exige madurez digital, inversión inicial y formación especializada, ofrece un retorno estratégico de largo plazo: capacidad institucional para convertir datos en decisiones informadas, reproducibles y socialmente relevantes.

7. Recomendaciones

El marco metodológico aquí propuesto constituye una herramienta estratégica para entidades que buscan estructurar, escalar y consolidar sus capacidades de analítica institucional, promoviendo la toma de decisiones basada en evidencia. A partir de los hallazgos teóricos y de la validación simulada, se plantean las siguientes recomendaciones clave:

7.1 Aplicación institucional del modelo

Se recomienda utilizar este modelo como **guía base para gobiernos, universidades y organizaciones no gubernamentales (ONGs)** que estén iniciando o formalizando procesos de análisis de datos a escala organizacional. Su enfoque modular permite iniciar desde cualquier fase según el nivel de madurez de cada institución, mientras que su integración con estándares internacionales (FAIR, DevOps, DataOps) facilita la interoperabilidad y la sostenibilidad del ecosistema.

Particularmente, este enfoque puede resultar valioso en áreas como:

- Políticas públicas basadas en datos.
- Evaluación de programas sociales y educativos.
- Monitoreo institucional en salud, medio ambiente o desarrollo territorial.
- Laboratorios de innovación pública o académica.

7.2 Líneas de investigación futura

A partir de la experiencia conceptual y simulada, se proponen tres líneas prioritarias para profundizar y extender el modelo en futuras investigaciones:

- **Adaptación sectorial del marco metodológico:** Estudiar cómo se comporta y ajusta el modelo en sectores específicos (salud, justicia, educación superior, agricultura), identificando particularidades de dominio, tipos de datos y desafíos de gobernanza.
- **Desarrollo de herramientas para evaluación de impacto de DataLabs:** Proponer y validar indicadores estandarizados para medir el valor

institucional, social y técnico de los proyectos DataLab, más allá de métricas puramente operativas.
- **Automatización de la gestión de tareas y despliegues:** Diseñar o adaptar plataformas que integren herramientas de orquestación (CI/CD), control de versiones, gestión de producto (Scrum) y visualización de avances, reduciendo la carga operativa de los equipos y mejorando la trazabilidad del proceso.

Anexos sugeridos:

- Plantilla editable de fases y checklist
- Diagrama de flujo de la metodología
- Ejemplo de backlog y sprint board
- Indicadores clave de madurez metodológica

Referencias:

- **Ambler, S. W., & Lines, M.** (2012). *Disciplined Agile Delivery: A practitioner's guide to agile software delivery in the enterprise*. IBM Press.
- **Arenas Contreras, D. A.** (2022). *Data science use cases in the manufacturing industry* [Tesis de maestría, University of St Andrews].
- **Atwal, H.** (2020). *Practical DataOps: Delivering agile data science at scale*. Apress.
- **Bergh, C., Davis, G., & Asay, D.** (2022). *DataOps Cookbook*. DataKitchen Press.
- **Brown, T.** (2009). *Change by design: How design thinking creates new alternatives for business and society*. Harvard Business Review Press.
- **Cervone, H. F.** (2011). Understanding agile project management methods using Scrum. *OCLC Systems & Services: International Digital Library Perspectives, 27*(1), 18–22.
- **Chapple, M., & Seidl, D.** (2020). *Cybersecurity and data science: Concepts, techniques, and tools*. Wiley.
- **Dehghani, Z.** (2022). *Data mesh: Delivering data-driven value at scale*. O'Reilly Media.
- **Demchenko, Y., Cuadrado-Gallego, J. J., & Chertov, O.** (2024). Data science projects management: DataOps, MLOps. En *Big Data Infrastructure Technologies for Data Analytics*. Springer.
- **Gopalan, A.** (2022). *The cloud data lake*. O'Reilly Media.
- **Kelleher, J. D., & Tierney, B.** (2018). *Data science*. MIT Press.
- **Kerzner, H.** (2021). *Project management: A systems approach to planning, scheduling, and controlling* (12ª ed.). Wiley.

- **Leffingwell, D.** (2018). *SAFe 4.5 reference guide: Scaled Agile Framework for Lean Enterprises*. Addison-Wesley.
- **Liedtka, J., Ogilvie, T., & Brozenske, R.** (2014). *Design thinking for the greater good: Innovation in the social sector*. Columbia Business School Publishing.
- **Packt.** (2023). *Data fabric architectures and governance*. Packt Publishing.
- **Piatetsky-Shapiro, G., Brachman, R., Khabaza, T., & Others.** (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS. https://www.the-modeling-agency.com/crisp-dm.pdf
- **Plattner, H., Meinel, C., & Leifer, L.** (2010). *Design thinking: Understand – improve – apply*. Springer.
- **Project Management Institute (PMI).** (2021). *A guide to the Project Management Body of Knowledge (PMBOK® Guide)* (7ª ed.). Project Management Institute. https://www.pmi.org
- **Reis, J., & Housley, M.** (2022). *Fundamentals of data engineering*. O'Reilly Media.
- **Warden, P.** (2018). *The data project manager*. Independently published.