



INTRODUCCIÓN A LA CIENCIA DE DATOS Y EL BIG DATA

CICLO DE VIDA DE UN PROYECTO DE CIENCIA
DE DATOS

Un proyecto está compuesto de procesos, etapas y un grupo de personas responsables de que esté se concluya de forma satisfactoria. Cabe mencionar que antes de hablar de un proyecto de Ciencia de Datos, era necesario definir los conceptos relacionados y los elementos involucrados. Por lo que ya debe de quedar claro que la Ciencia de Datos es una disciplina que a través de la manipulación de grandes volúmenes de datos (Big Data) y el análisis utilizando herramientas tecnológicas o modelos matemáticos es posible saber que ha ocurrido, que va a ocurrir o que se podría mejorar (análisis descriptivo; predictivo o prescriptivo) con respecto a los datos.



Ahora bien, un proyecto de Ciencia de Datos toma sus bases en la forma de administrar y ejecutar un proyecto tecnológico, específicamente hablando de un proyecto de software. Un proyecto de software tiene un ciclo, el cual es el proceso que se sigue para construir, entregar y hacer evolucionar el software, desde la concepción de una idea hasta la entrega y retiro del sistema. Se definen las distintas fases intermedias que se requieren para validar el desarrollo de un software, es decir, para garantizar que el software cumpla los requisitos para la aplicación y verificación de los procedimientos de desarrollo, se asegura de que los métodos utilizados son apropiados. Y a partir de este tipo de proyectos, es que surge el ciclo de vida de un proyecto de Ciencia de Datos.

Identificar qué etapas debemos seguir a la hora de abordar un proyecto de Ciencia de Datos es fundamental para estructurar y analizar que recursos son necesarios y en qué fase tendrán más relevancia su implicación. De este modo estaremos más preparados ante cualquier eventualidad, pudiendo estimar los esfuerzos que debemos asumir, así como identificar la viabilidad y reducción de los costos operativos del proyecto. La identificación de estas etapas permitirá a los directivos y gerente del proyecto comprender el alcance del mismo, vigilar y tratar los riesgos inherentes a esta nueva tecnología y aclarar y resolver los problemas que surjan a lo largo del desarrollo del proyecto relacionados con su gestión. Cada fase de un proyecto de Ciencia de Datos contendrá sus propias tareas y demandas de realización que serán más o menos críticas atendiendo a la madurez y grado de conocimiento, habilidad y/o limitación de recursos en el que la empresa aborda dichos proyectos, así como el tiempo de adopción del uso y/o utilización del servicio/producto por parte del mercado.

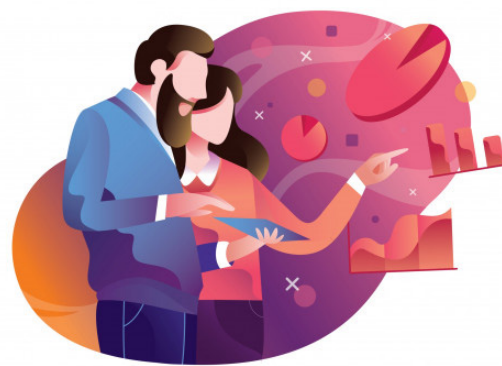
Por otro lado, debemos tener en cuenta que la Ciencia de Datos se identifica dentro de un contexto multidisciplinar, donde coexisten diferentes perfiles profesionales que cubren una función determinada. En este sentido, debemos aclarar que cada proyecto de Ciencia de Datos deberá asumir sus propios requerimientos que definirán sus requisitos y alcance.

Del mismo modo, cada infraestructura o software de terceros que se emplee para el desarrollo de los modelos y puesta en marcha del proyecto posee sus propias exigencias que fijarán, de una forma u otra, la toma de decisiones en cuanto a que recursos son necesarios y que etapas y tiempos son los necesarios. Al mismo tiempo, cada problema a solventar, cada producto a desarrollar tiene sus propias necesidades de negocio y por tanto determinarán el flujo y seguimiento de las fases que deben de aplicarse a lo largo del ciclo de vida de un proyecto de Ciencia de Datos.



En todos los sistemas para la toma de decisiones que coexisten en la actualidad se vuelven indispensables las técnicas y tecnologías que engloban Big Data para la transmisión, almacenamiento y análisis de grandes volúmenes de datos. Así como la aplicación de una analítica avanzada que permita predecir situaciones futuras y seleccionar las estrategias y toma de decisiones más optima. Por ello, las técnicas que engloban la Ciencia de Datos, entre las que destacan el Data Mining (Análisis Descriptivo) y el Aprendizaje Automático como Machine Learning (ML) (Análisis Predictivo) y Deep Learning (DL) (Análisis Prescriptivo) son las más utilizadas hoy en día, ya que posibilitan enriquecer y crear conocimiento de negocio, optimizando de esta manera la toma de decisiones y extrayendo nueva información oculta entre los datos generados por la empresa.

Podemos describir estos sistemas como una evolución de los denominados sistemas KDD (Knowledge Discovery in Databases) (Figura 1). En diferentes estudios realizados a lo largo de esta última década se confirma que más del 70% de los esfuerzos dentro de un proyecto de Ciencia de Datos se centra el tratamiento de los datos para obtener el conocimiento necesario para la toma de decisiones y su aportación al negocio.



Hay que enfatizar que la Ciencia de Datos no hace referencia únicamente en tratar y gestionar grandes volúmenes de datos de forma ordenada, hecho que si los diferencia, de los sistemas KDD, sino también es necesario atender a su variabilidad, calidad y representatividad de los mismos, creando un producto de datos final. Este producto de datos pueden ser un cuadro de mandos, un recomendador, clasificador o cualquier respuesta que facilite la toma de decisiones y actuaciones.

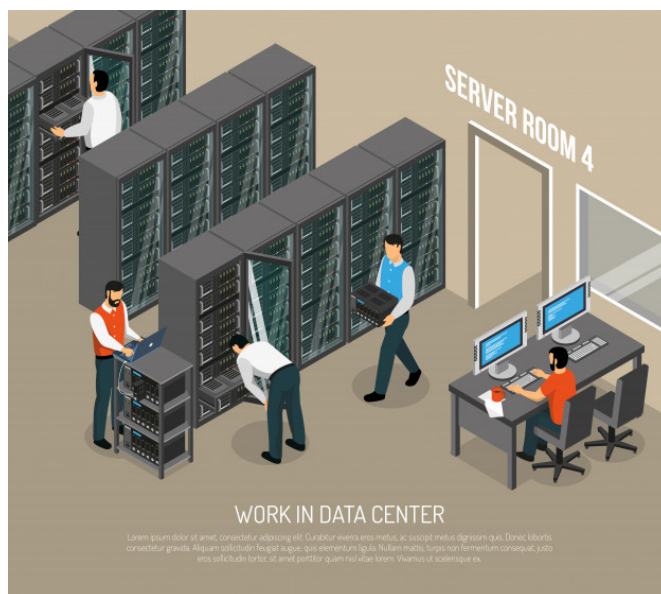
Para ello es necesario mantener los procesos de disposición de los datos, clasificación, selección, limpieza y reducción de los mismos. Así como, la introducción de conocimiento previo, y la misma interpretación de los resultados. Sin ello no sería posible alcanzar una fiabilidad y eficacia recomendable en la toma de decisiones. Hecho que se asemeja a como un ser humano aprende obteniendo información del entorno: captamos información del entorno, almacenamos dicha información que nos llega por diferentes fuentes (sentidos), la filtramos, nos quedamos con aquella información que nos interesa y sobre ella razonamos y tomamos decisiones.

En este último paso de razonamiento aplicamos la experiencia adquirida, el propósito o intención y las emociones, para discriminar que actuaciones pueden resultar menos beneficiosas atendiendo a los errores cometidos en situaciones pasadas y la información obtenida por diferentes fuentes. Aspectos que vienen a tenerse en cuenta en el ciclo de vida de un proyecto de Ciencia de Datos ya sea con técnicas de Data Mining o con las técnicas de Machine Learning y más recientemente, Deep Learning.

Al igual que ocurre con el desarrollo de software donde existe una Ingeniería del Software con diferentes metodologías que permiten optimizar los tiempos y mejora de la calidad del mismo, facilitando de este modo la interconexión entre todos los miembros del equipo de proyecto. Los desarrollos y proyectos de Ciencia de Datos deben construirse alrededor de un equipo multidisciplinar que debe trabajar coordinado, comunicado e integrado de forma ágil con todos los departamentos presentes en la cadena de valor del proyecto y con una cultura centrada en el dato y la generación de conocimiento útil, con una fuerte estandarización y automatización de los procesos que permita la escalabilidad del proyecto por complejo que sea.

Es posible identificar este tipo de proyectos como una combinación entre ciencia e ingeniería, donde los conocimientos estadísticos, matemáticos avanzados e investigación aplicada son necesarios pero donde también es imprescindible, el conocimiento de los algoritmos, habilidades en el tratamiento y análisis de datos heterogéneos, gestión del software, organización y diligencia para adelantarse y solventar los posibles

riesgos, adaptación de los datos, algoritmos e infraestructuras al negocio y los objetivos marcados y por supuesto, su implantación y puesta en producción, evaluando su fiabilidad y su cumplimiento tanto en las restricciones éticas y normativas como en las necesidades y perspectivas de los clientes.



Hoy en día existen diferentes metodologías de actuación para la gestión y optimización de este ciclo de vida. Uno de ellos es la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) creado en 1999 por SPSS, NCR y DaimlerChrysler, el cual mantiene un proceso estándar en seis fases que fue concebido para el desarrollo de proyectos de Data Mining (Minería de Datos), en donde se busca la recolección y análisis de grandes volúmenes de datos. Otro estándar utilizado para proyectos de Data Mining basado en herramientas comerciales es el modelo SEMMA (Sample, Explore, Modify, Model, Assess) el cual se basan en el estándar CRISP-DM y creada por SAS Institute en 1998. Por lo general y dado que los proyectos de Data Mining son sensibles y deficitarios de las fuentes generadoras de datos, un modelo que se exterioriza de los proyectos de CRM hacia los proyectos de Ciencia de Datos es el modelo Catalyst creado en 2003. Una de las diferencias más importantes entre estos tipos

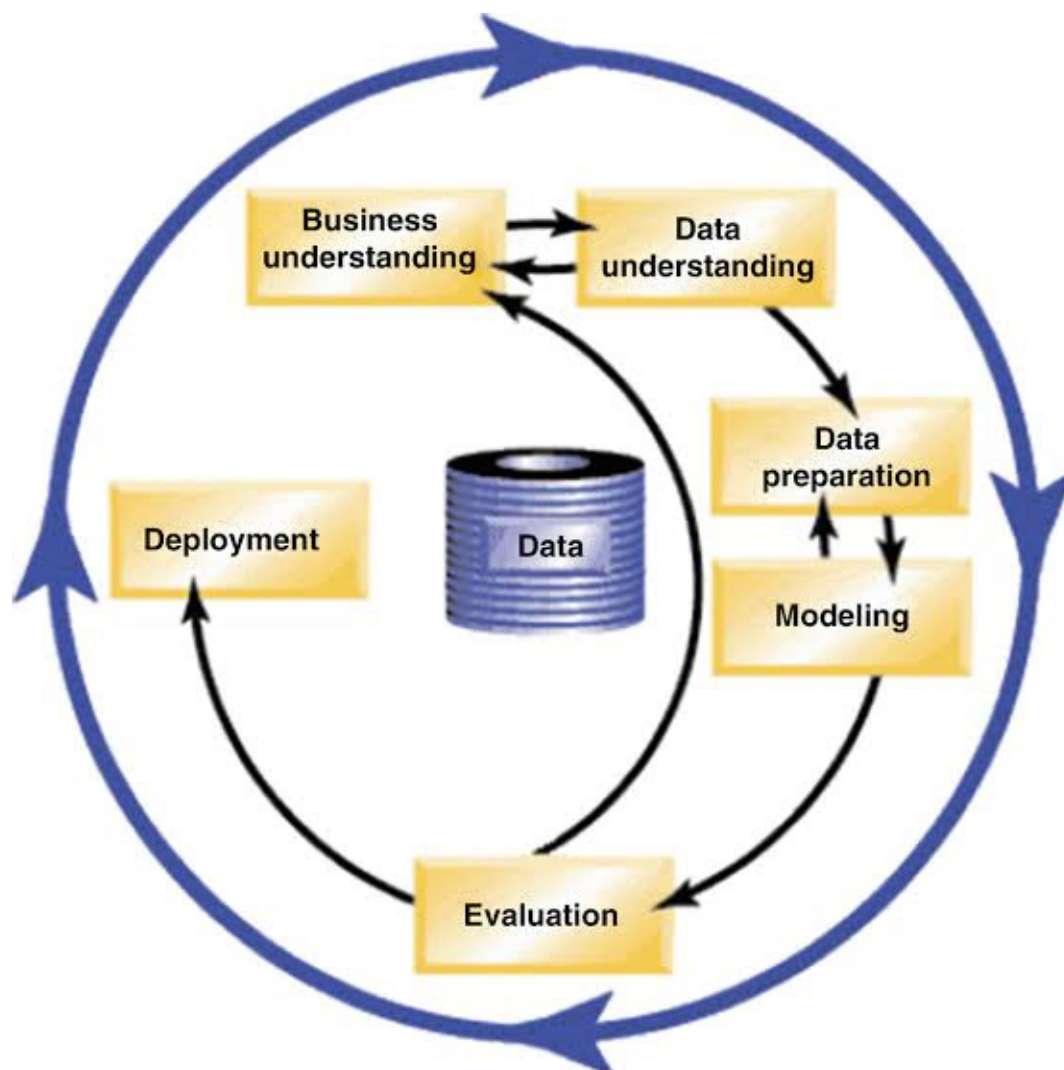
de metodologías es que mientras CRISP-DM, KDD o Catalyst se centran en las necesidades y comprensión del negocio, la metodología SEMA está más orientada al empleo de estadísticos para el muestreo de los datos.



En todas estas metodologías se enfatiza en la fase de identificación de fuentes de datos y la preparación y procesamiento de los mismos, así como la necesidad de evaluar el algoritmo de extracción de patrones de

conocimiento acorde a los datos que manejamos y los objetivos marcados. Por otra parte, la evaluación de los resultados en CRIPS-DM se realiza en base al desempeño del modelo elegido y los objetivos marcados, mientras que en la metodología SEMA sólo hace referencia al desempeño del modelo. En el caso de Catalyst la evaluación de los resultados se realiza únicamente en relación a los objetivos y requerimientos de la estrategia de negocio. Un aspecto a tener en cuenta es que la metodología SEMA está orientada a las herramientas de SAS y por tanto los algoritmos y modelos que SAS proporciona, mientras que con las demás metodologías el analista de Ciencia de Datos puede emplear las herramientas y modelos que desee. Entre las diferentes metodologías descritas es la de CRISP-DM la más utilizada en el mercado, existen diferentes fabricantes que aportan herramientas para el seguimiento del proyecto utilizando esta metodología, existiendo una red mundial de utilización de esta metodología. Entre los fabricantes más destacados en esta red se encuentran como Teradata, Sgi, DESPSS, IBM, OHRA y consultoras de prestigio como Deloitte, ICL, ABB, etc.

La metodología CRISP-DM (Cross Industry Standard Process) utilizada en data mining se presenta con las siguientes fases:



1. **Comprensión del negocio:** En esta fase se identifican los objetivos a conseguir después de un estudio pormenorizado del negocio, exigencias y necesidades del cliente. Crea un plan estratégico para alcanzar dichos objetivos con unos requerimientos de fiabilidad y calidad mínimos. Se debe tener en cuenta la regularización y normativas de ciberseguridad y privacidad de datos y sistemas informáticos.
2. **Adquisición de los datos:** Identificar los datos necesarios para la consecución de los objetivos. Reconocer las fuentes de datos. Describir los tipos de datos con los que vamos a trabajar e identificar aquellos que realmente son necesarios. Reconocer problemas en la calidad de los mismos, como por ejemplo si existen datos repetidos, incompletos, inconsistentes, con errores, entre otros.
3. **Preparación de los datos:** Procesar los flujos de datos, solventar problemas de datos faltantes, controlar las inconsistencias de los flujos de datos y realizar la limpieza y estandarización de los datos, generación de variables, integración de diferentes conjuntos de datos, etc.
4. **Modelación:** Determinar qué modelo o técnica es el más apropiado para la resolución del problema a tratar y que técnicas a aplicar de forma consistente atendiendo a los datos que tenemos, los recursos y necesidades. Por lo general, se puede volver a la fase anterior para trabajar con los datos y tener una entrada de los mismos, acorde a las necesidades del modelo. En esta fase se debe crear los test de evaluación y desempeño del sistema para estudiar la calidad y fiabilidad de los resultados obtenidos con el modelo seleccionado y los objetivos marcados.
5. **Evaluación e Interpretación:** Visualización y análisis de los datos obtenidos y su correspondencia sobre los objetivos, la fiabilidad y calidad deseada.
6. **Despliegue del Modelo:** Se visualiza el conocimiento y los resultados obtenidos y se muestran al cliente.
7. **Operaciones:** Realizar las acciones que el cliente vea pertinentes acorde a los resultados obtenidos. Además, pasamos a una fase de seguimiento y mantenimiento del modelo acorde por ejemplo al periodo de validez de los resultados o modelos utilizados, así como los objetivos de negocio que pueden variar con el tiempo. Puede ocurrir que la fiabilidad de los resultados del modelo baje por lo que se debe retomar el proyecto desde el principio.

Nota importante: Un aspecto relevante a destacar y que no se prevé en las metodologías indicadas es el estudio del impacto sobre el cliente con respecto a la incorporación e integración de estos modelos y sistemas de flujo de información en sus procesos internos.

Podemos encontrarnos otra metodología centrada en desarrollos en nube (Cloud) la cual es ampliamente utilizada por las grandes compañías que ofrecen servicios de IA y aprendizaje Automático. Una de las grandes diferencias entre la Ingeniería de Software tradicional y los proyectos de Ciencia de Datos, es que mientras la primera se centra en un desarrollo determinista de código, los

desarrollos de los proyectos de Ciencia de Datos basan en la creación de modelos para el tratamiento y análisis de datos para la generación de conocimiento. Estos modelos pueden perder su fiabilidad con el tiempo y por tanto debe existir un seguimiento y mantenimiento del mismo, y de los flujos de datos que nutren dicho modelo. Es en este punto donde se hace énfasis la nueva metodología denominada ModelOps, el cual se encarga de automatizar el despliegue del modelo, así como de su seguimiento, supervisión y mantenimiento tal que acelera el desarrollo y mejora la escalabilidad de este tipo de proyectos.

ModelOps se basa en la metodología Devops la cual se emplea para el desarrollo de aplicaciones, mientras que ModelOps se centra en acelerar el proceso de creación de modelos desde su fase inicial de laboratorio, validación y pruebas hasta su despliegue con la calidad y fiabilidad esperada acorde a los objetivos establecidos. Por otra parte, el fuerte auge que está teniendo la Inteligencia Artificial (IA) y el Aprendizaje Automático (Machine Learning) han hecho que las grandes compañías como Microsoft, Google, Amazon o IBM creen un conjunto de banco de modelos y herramientas para el desarrollo de estos proyectos de aprendizaje automático de forma ágil y basadas en tecnología en la nube. Este nuevo conjunto de servicios, banco de modelos junto a la metodología ModelOps, permite el desarrollo y gestión de forma ágil de este tipo de proyectos. Facilitando la democratización y acercamiento de estas tecnologías a todos los niveles de una organización, por ejemplo, dado su alto grado de abstracción y fácil uso.



La alta flexibilidad de esta metodología basada en estas herramientas y servicios en nube permite construir todo el ciclo de vida de un proyecto de Ciencia de Datos con unos pocos clics. Se centra en el concepto de tubería (pipeline), es decir, el usuario une con un conector las tareas que debe realizar en cada paso del proceso (flujos de trabajo). Por ejemplo, permite incluir cualquier control de calidad que el analista de Ciencia de Datos desee (Control seguridad, sesgo, variabilidad, verificaciones de cumplimiento, etc..). Así como hacer un seguimiento de los datos y los modelos a lo largo de todo el ciclo de vida del proyecto. Además, permite de forma automática mejorar los modelos mediante un bucle retroalimentado desde el interfaz de usuario al modelo de backend. Por lo general estas herramientas soportan una interfaz de usuario donde pueden construir un grafo acíclico en el cual, cada nodo representa la tarea a realizar y cada borde define el flujo de control. Además, se basa en un patrón de eventos, los cuales permiten controlar el tiempo de ejecución de los modelos y las aplicaciones.

Una de las ventajas de esta metodología es la posibilidad de gestionar y mantener versiones distintas de un mismo modelo (entrenados con diferentes conjuntos de datos o valores distintos de sus parámetros de configuración). Este tipo de servicios y herramientas basadas en metodología ModelOps, por lo general, proporcionan una plataforma integrada en nube que permite a los usuarios administrar e implementar modelos usando un flujo de trabajo colaborativo y automatizado. Y que pueden utilizarse tanto en entornos de producción en la empresa como en entornos de investigación. Algunos ejemplos de plataformas son los siguientes:

- Watson Machine Learning (WML): <https://www.ibm.com/es-es/cloud/machine-learning>
- Azure Machine Learning (AML): <https://azure.microsoft.com/es-es/free/machine-learning/>
- AWS Machine Learning: <https://aws.amazon.com/es/machine-learning/>
- Cloud Machine Learning: <https://cloud.google.com/products/ai/>

Un ciclo de vida particular para trabajar con esta metodología ModelOps es la que desarrolló Microsoft y que se puede visualizar en la figura 4.

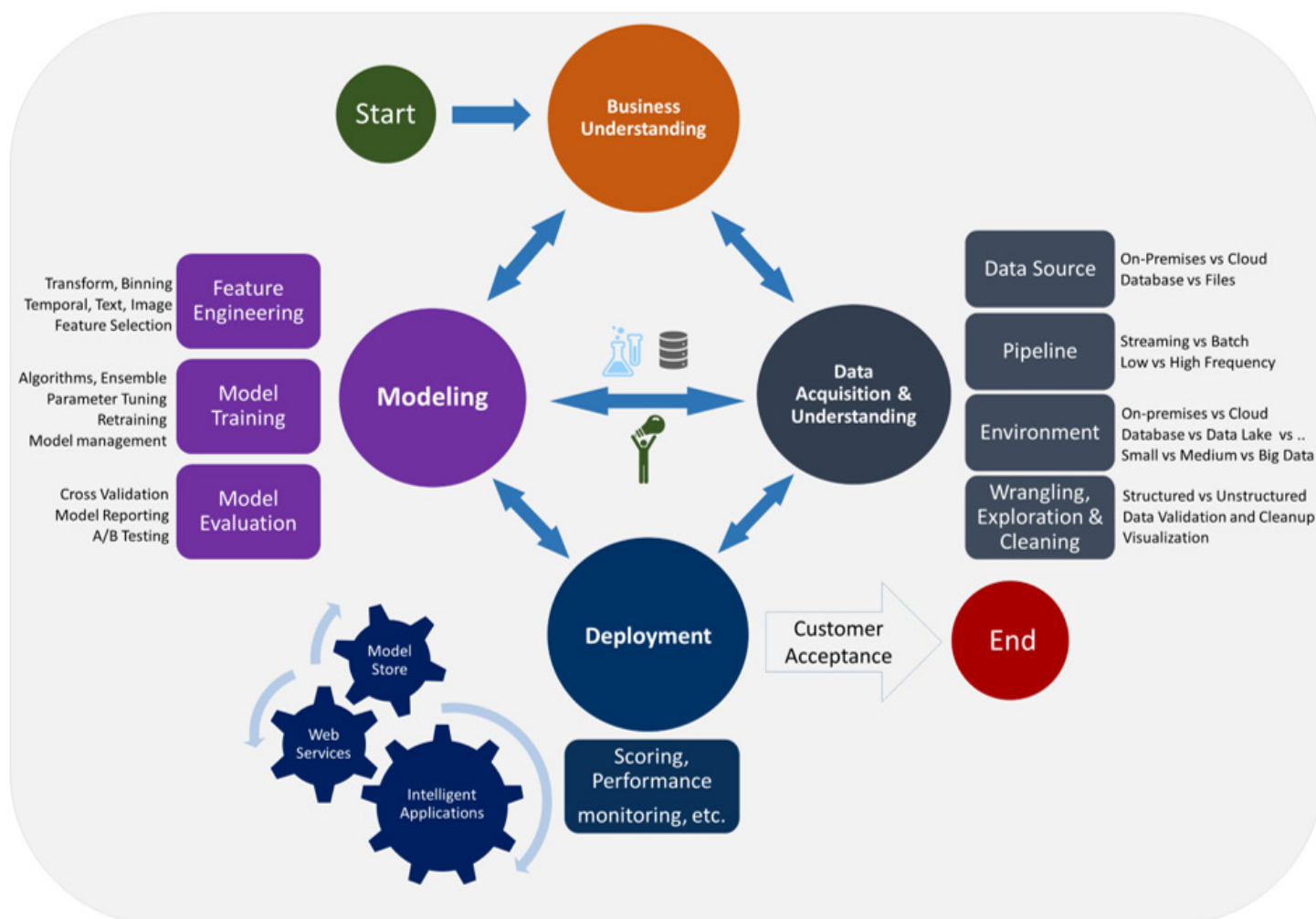



Figura 4. Ciclo de vida de un proyecto de Ciencia de Datos desarrollado por Microsoft.



En este ciclo de vida podemos observar las mismas fases que encontrábamos con la metodología CRISP-DM como la comprensión del negocio, la adquisición de los datos o la fase de modelado, pero no de forma iterativa. La escalabilidad y flexibilidad de esta metodología y plataforma permite trabajar en paralelo diferentes tareas necesarias del ciclo de vida del proyecto.

Al margen de estos diferentes modelos, hay que tener en cuenta que, por un lado, en todo proyecto y más en especial en los proyectos de Ciencia de Datos existen diferentes retrasos reiterados. Por lo general, el trabajar con datos y fuentes casi siempre desconocidas hace crecer la incertidumbre y los riesgos asociados, existiendo la posibilidad de retrasos por no conocer el formato real de los datos, desconocer el tipo de fuente o simplemente no caer en que los datos no están totalmente formateados en relación con las tareas que se ejecutan posteriormente. Esto hace que en las fases del ciclo de vida de un proyecto de Ciencia de Datos no se pueda contemplar como algo lineal, siendo altamente iterativos y cíclicos existiendo grandes dependencias entre el equipo de Ciencia de datos y los demás equipos involucrados en el proyecto.

Se prohíbe la reproducción total o parcial de esta obra por cualquier medio sin previo y expreso consentimiento por escrito del Instituto Tecnológico y de Estudios Superiores de Monterrey.

D.R. © Instituto Tecnológico y de Estudios Superiores de Monterrey, México. 2019 Ave. Eugenio Garza Sada 2501 Sur Col. Tecnológico C.P. 64849 Monterrey, Nuevo León | México



**Tecnológico
de Monterrey**