Programa de formación

MACHINE LEARNING AND DATA SCIENCE MLDS

Facultad de INGENIERÍA

UNIVERSIDAD NACIONAL DE COLOMBIA

Módulo 6

Metodologías de Desarrollo de Aplicaciones con Machine Learning

Unidad 3

Ciclo de Vida de Ciencia de Datos

Facultad de INGENIERÍA

UNIVERSIDAD NACIONAL DE COLOMBIA

Bienvenida

Jorge Eliécer Camargo Mendoza, PhD.

UNIVERSIDAD NACIONAL DE COLOMBIA

https://dis.unal.edu.co/~jecamargom/

jecamargom@unal.edu.co

Departamento de Ingeniería de Sistemas e Industrial
Facultad de Ingeniería
Universidad Nacional de Colombia
Sede Bogotá

Tabla de contenidos

Objetivos de aprendizaje

Unidad 3 – Ciclo de Vida de Ciencia de Datos

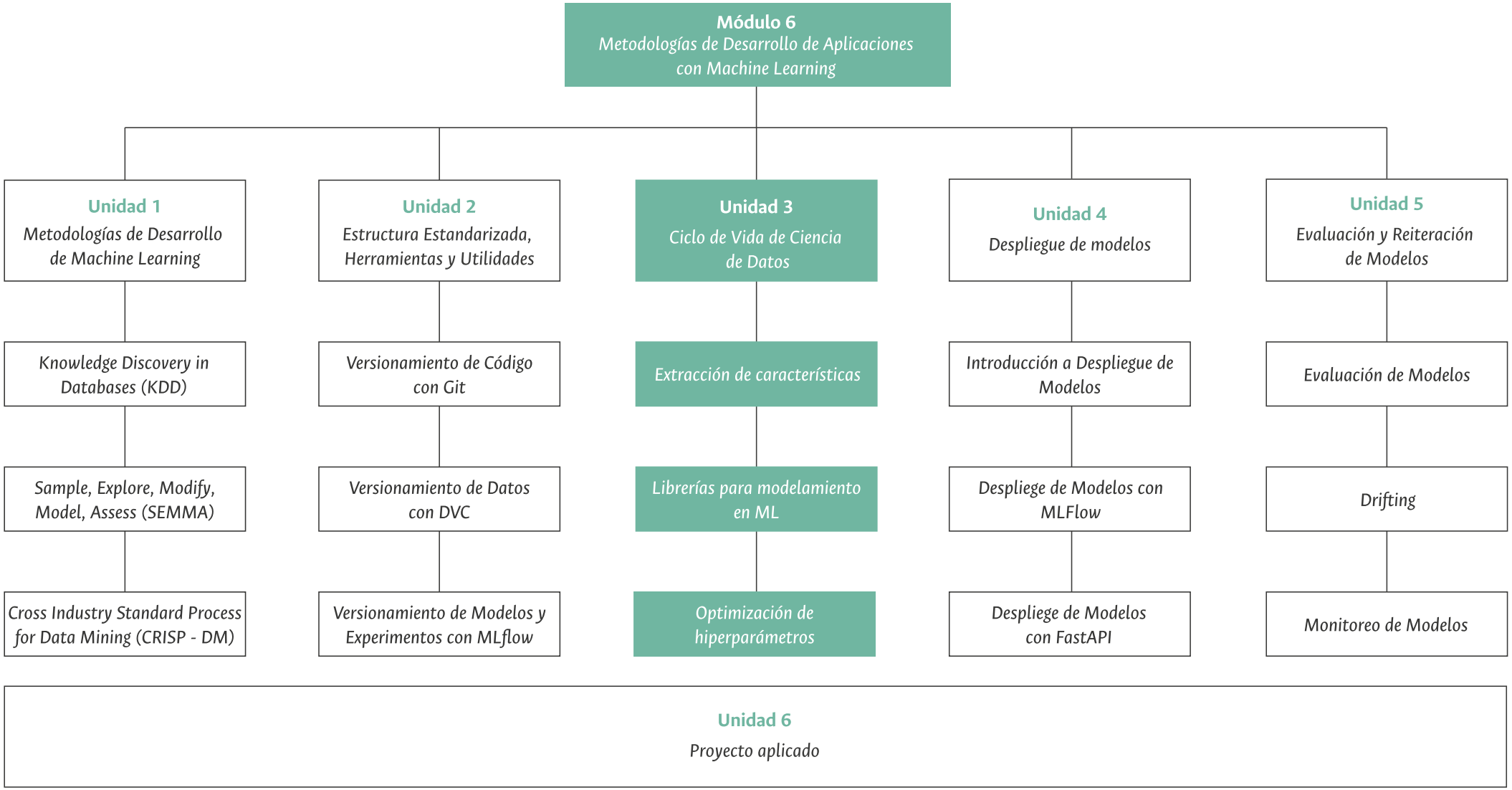
Al finalizar la unidad usted deberá ser capaz de:

Describir el detalle de cada uno de los componentes del ciclo de vida de ciencia de datos.

Seleccionar la herramienta más apropiada para la extracción de características según el tipo de datos.

Usar herramientas modernas para modelamiento, comparación y selección de modelos.

Mapa de contenidos de la unidad



Team Data Science Process

PRE - SPRINT

Se lleva a cabo el
Entendimiento del negocio

Se realiza el
Entendimiento de los datos

SPRINT 0

Se plantea
El diseño de la solución

Se crea la
Planeación de la arquitectura del sistema

Se realiza la
Preparación avanzada de los datos

Se desarrolla la
Planeación del modelado

SPRINT 1 - N

Se hace
Ingeniería de datos

Se efectúa la
Preparación de los datos

Se desarrolla el
Modelado del sistema

OPERACIONAL

Se lleva a cabo la
Evaluación de los resultados

Se realizan
Despliegues eventuales

Se hacen
Iteraciones continuas

En octubre de 2016, Microsoft propuso una nueva metodología de ciencia de datos llamada Team Data Science Process (TDSP). Esta metodología se puede ver como una combinación entre SCRUM y CRISP-DM.

El objetivo es guíar a los equipos de ciencia de datos con un enfoque sistemático y estructurar proyectos de esta área en aspectos como: **colaboración en equipo**, **formación**, **calidad** y **eficiencia**.

Team Data Science Process

Componentes

TDSP tiene 4 componentes fundamentales:

Ciclo de vida de ciencia de datos ⚠️

Se describen los pasos completos que siguen los proyectos de ciencia de datos que hacen parte de aplicaciones inteligentes. Estas aplicaciones implementan modelos de *machine learning* o IA para realizar un análisis predictivo.

Estructura estandarizada de proyecto

Es un conjunto de elementos y prácticas comunes que se utilizan para planificar, ejecutar y controlar proyectos. Esta estructura está diseñada para asegurar que los proyectos se lleven a cabo de manera consistente y efectiva

Herramientas y utilidades

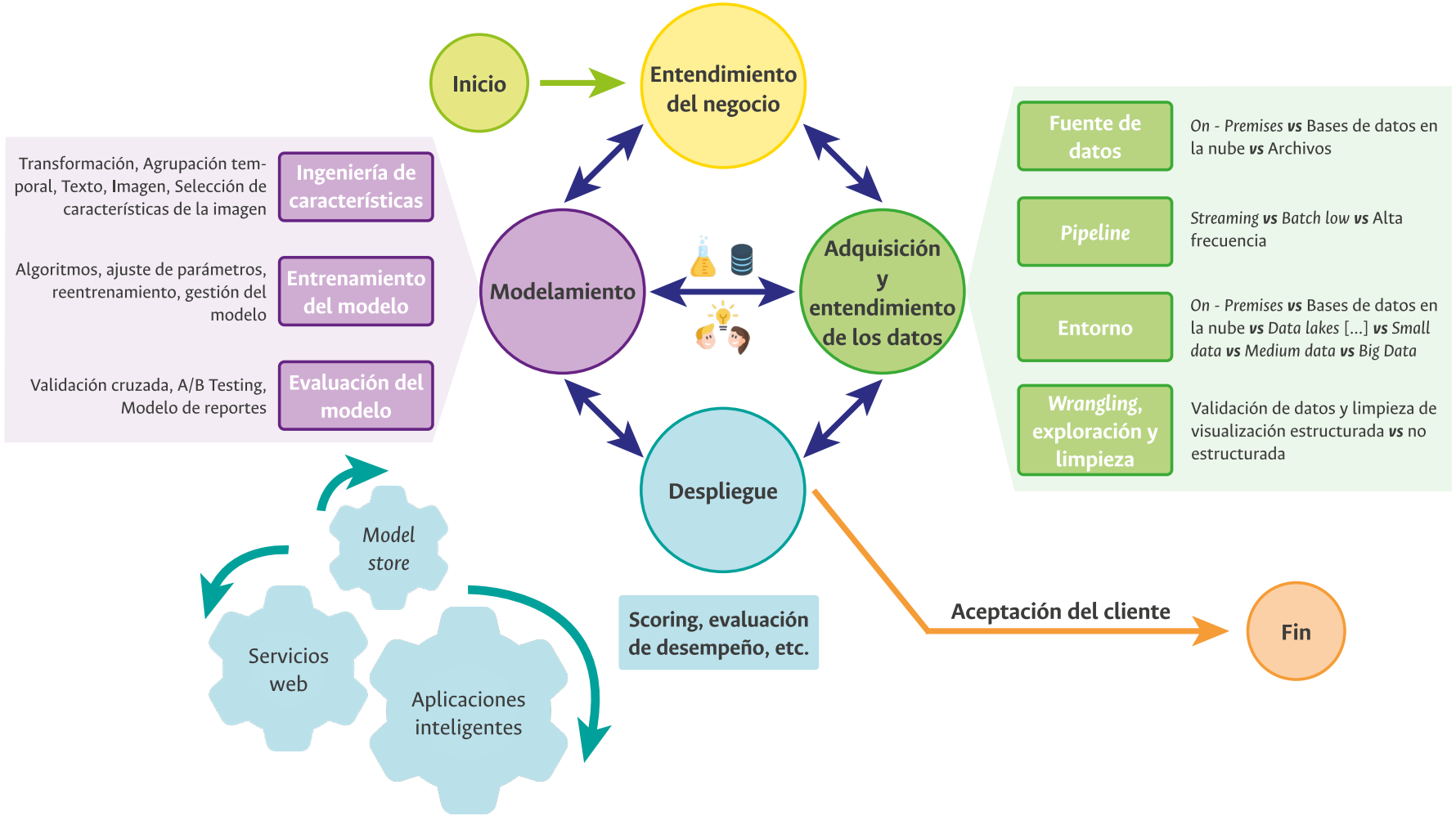
Programas y recursos utilizados para desarrollar, gestionar y mantener el software durante todo su ciclo de vida. Estas herramientas pueden incluir software de desarrollo, de gestión de proyectos, de control de versiones, de prueba y depuración, de documentación, etc.

Infraestructura y recursos

La infraestructura se refieren a los elementos necesarios para llevar a cabo un proyecto que fueron determinados en el proceso de planeación. Los recursos pueden desde hardware y software hasta recursos humanos y financieros.

Ciclo de vida de Ciencia de Datos

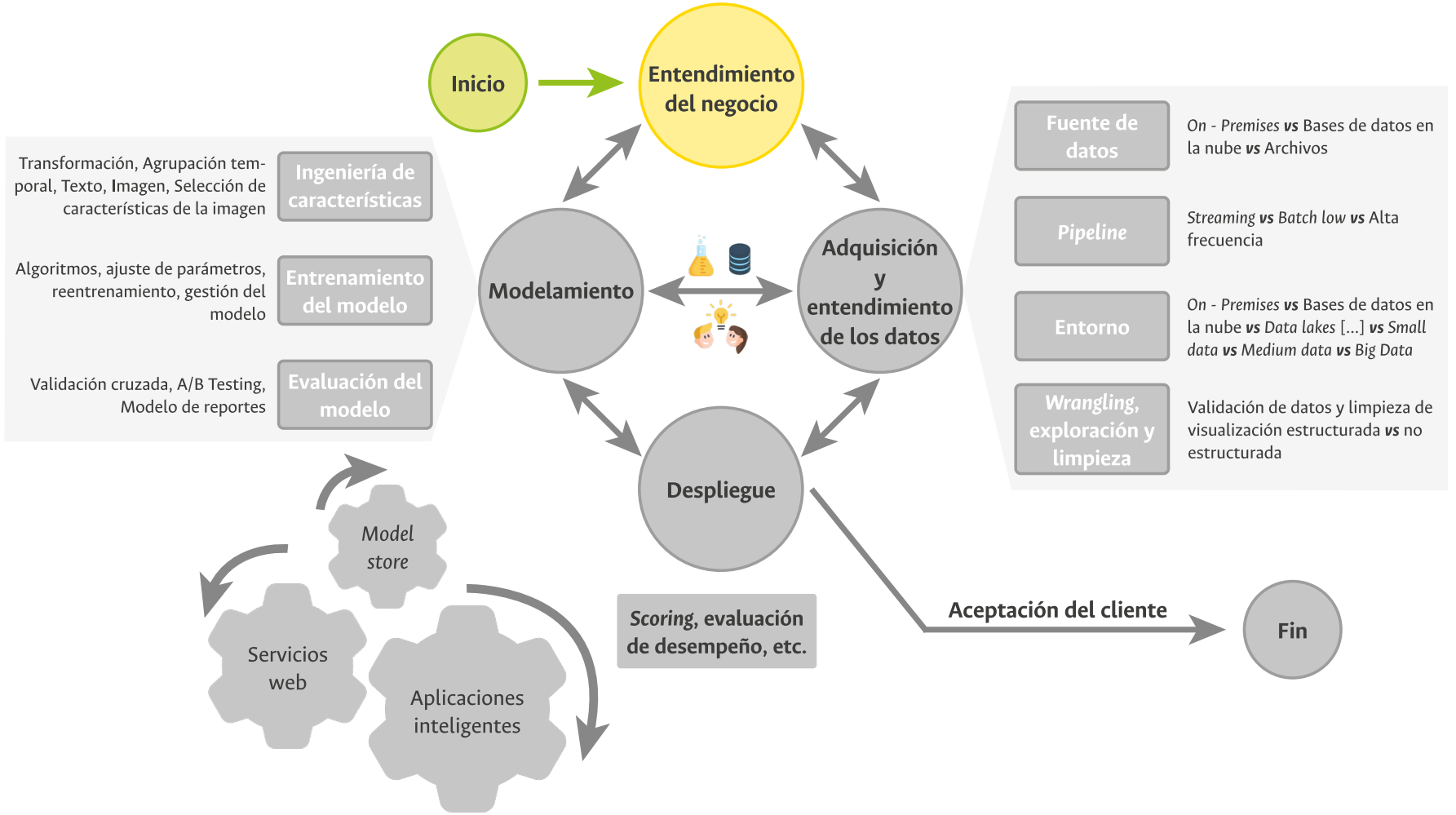
Busca estructurar el desarrollo de proyectos de ciencia de datos y tiene una gran similitud con las metodologías clásicas de ciencia de datos (KDD, SEMMA, CRISP-DM). Aunque se define específicamente para TDSP, puede ser reemplazado por cualquier otro ciclo de vida.



Ciclo de vida de Ciencia de Datos

Entendimiento del negocio

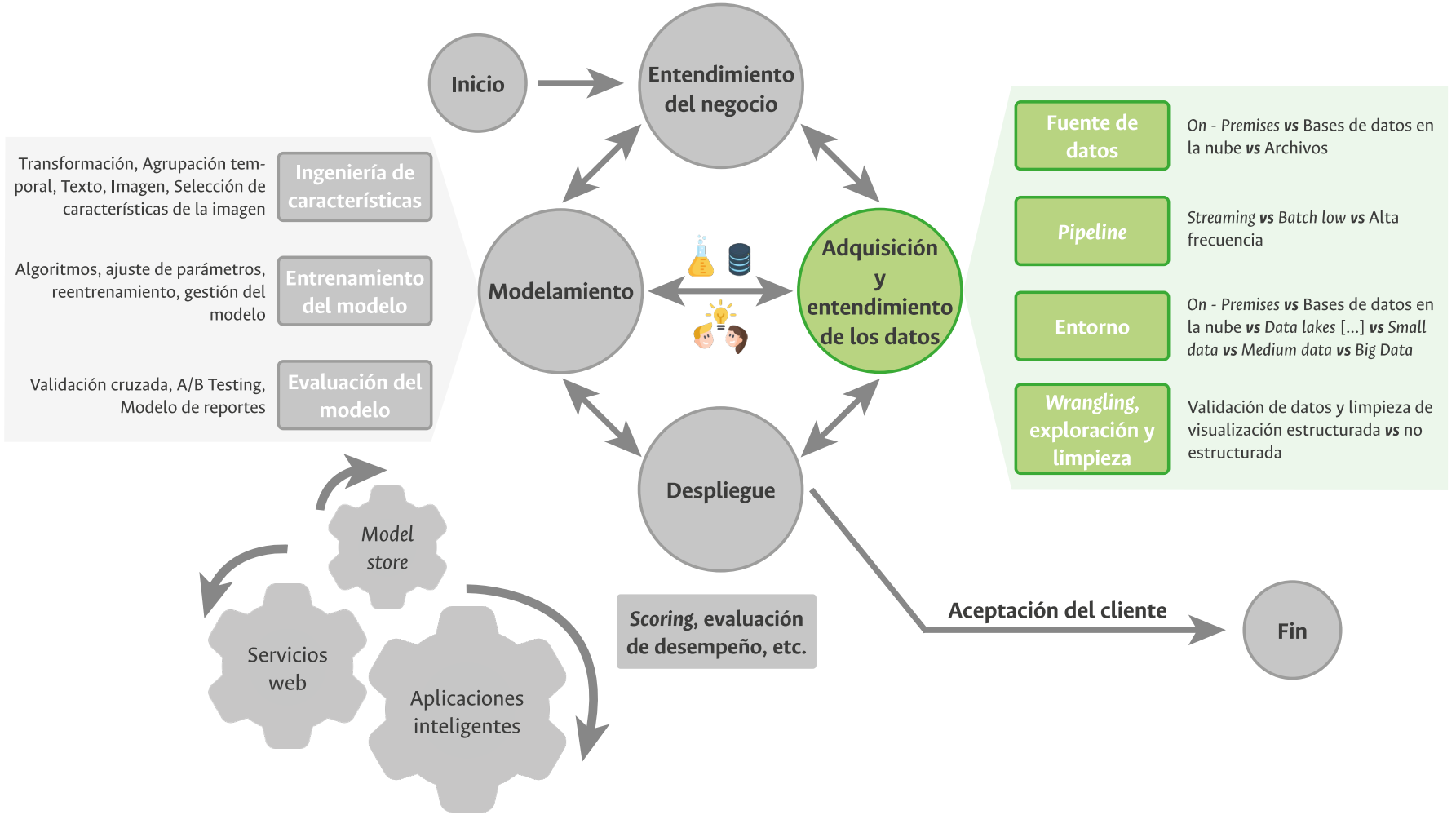
Se refiere a la comprensión de objetivos y requisitos. Permite a los desarrolladores y miembros del equipo de proyecto entender las necesidades y requerimientos del negocio y diseñar soluciones que se ajusten a las mismas.



Ciclo de vida de Ciencia de Datos

Adquisición y entendimiento de los datos

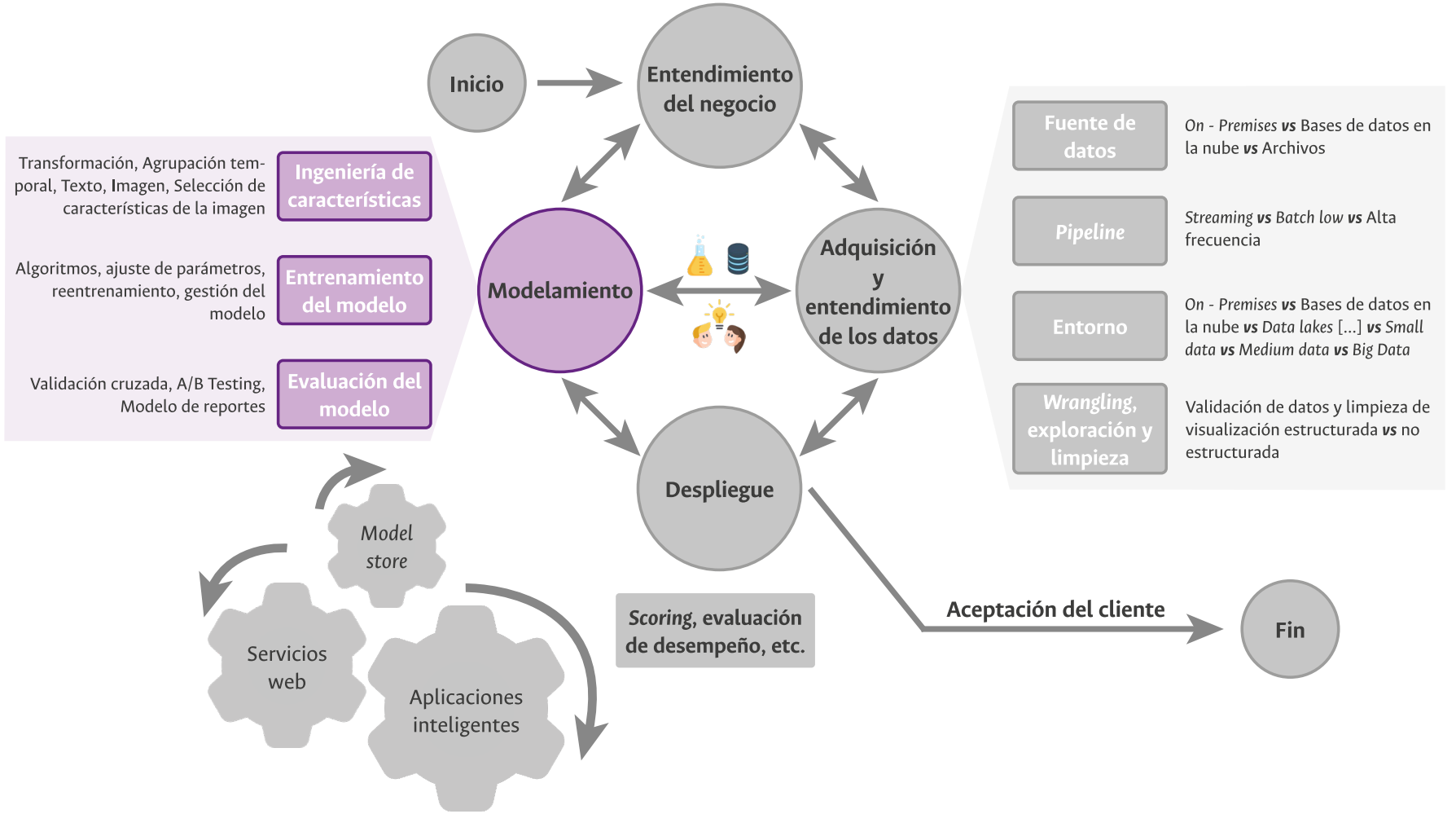
Se refiere a la comprensión de objetivos y requisitos. Permite a los desarrolladores y miembros del equipo de proyecto entender las necesidades y requerimientos del negocio y diseñar soluciones que se ajusten a las mismas.



Ciclo de vida de Ciencia de Datos

Modelamiento

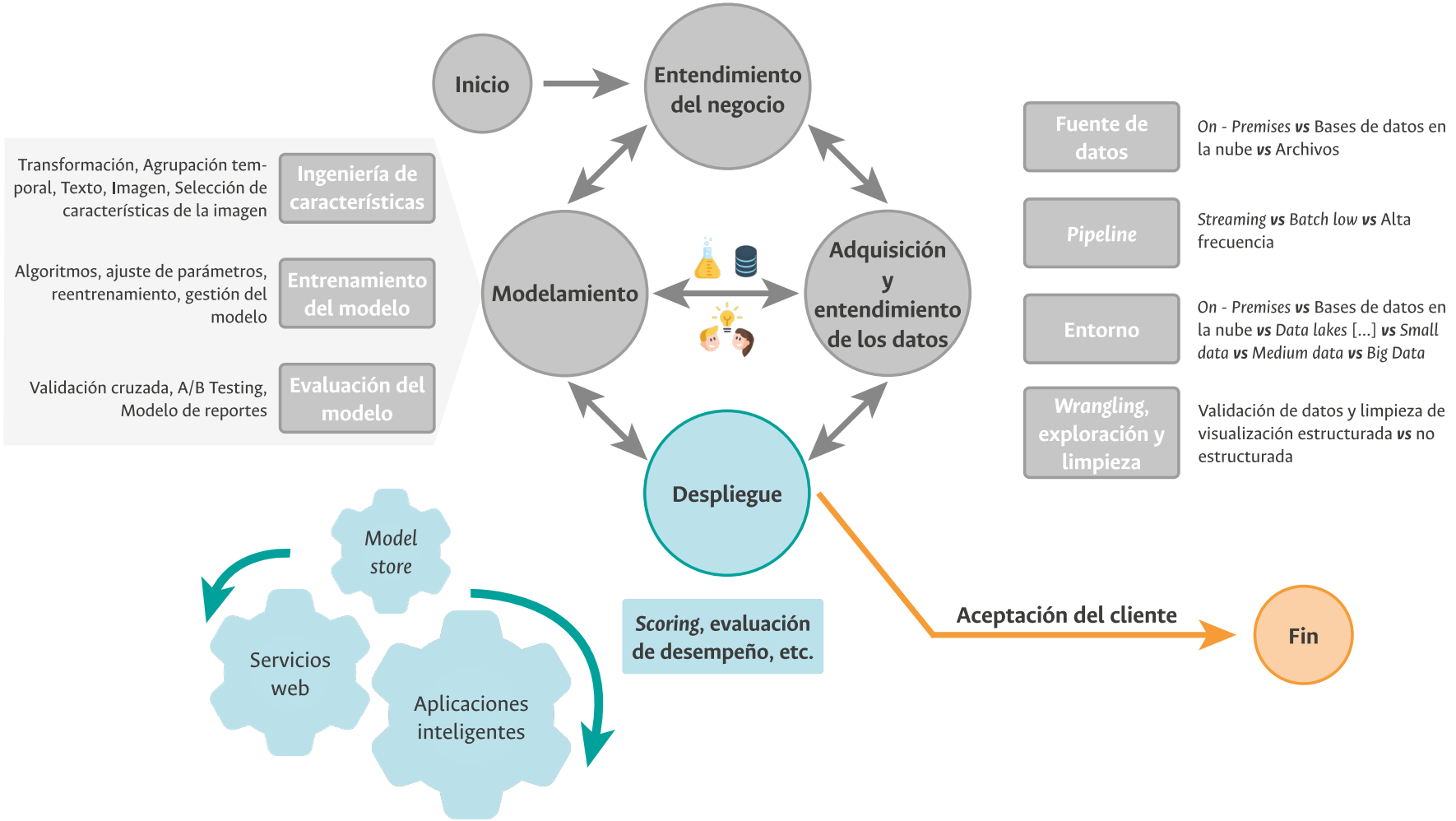
En esta etapa, se construyen modelos matemáticos o estadísticos que pueden predecir, clasificar o agrupar datos. Se selecciona el modelo adecuado para resolver el problema y se entrena el mismo utilizando los datos recopilados.



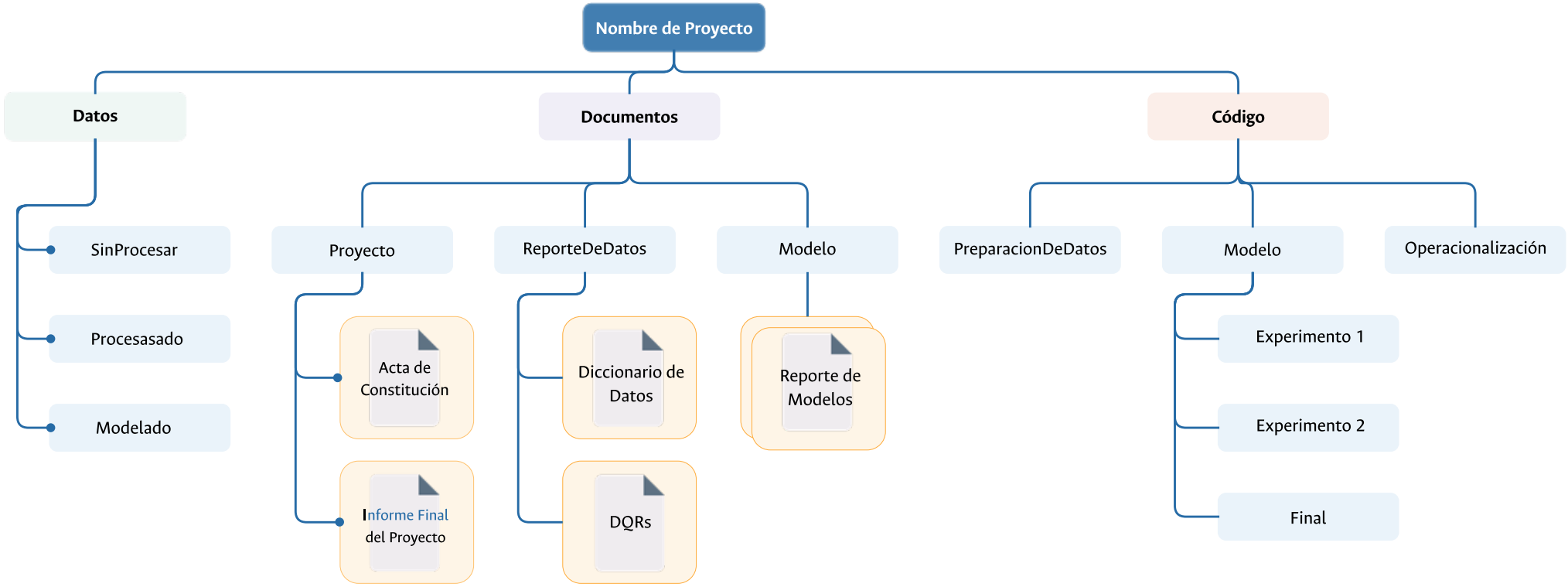
Ciclo de vida de Ciencia de Datos

Despliegue

El modelo o sistema de aprendizaje automático desarrollado es implementado en producción y puesto a disposición de los usuarios finales. Es en esta fase en la que se lleva a cabo la implementación del modelo en un ambiente de producción.



Estructura Estandarizada de Proyecto

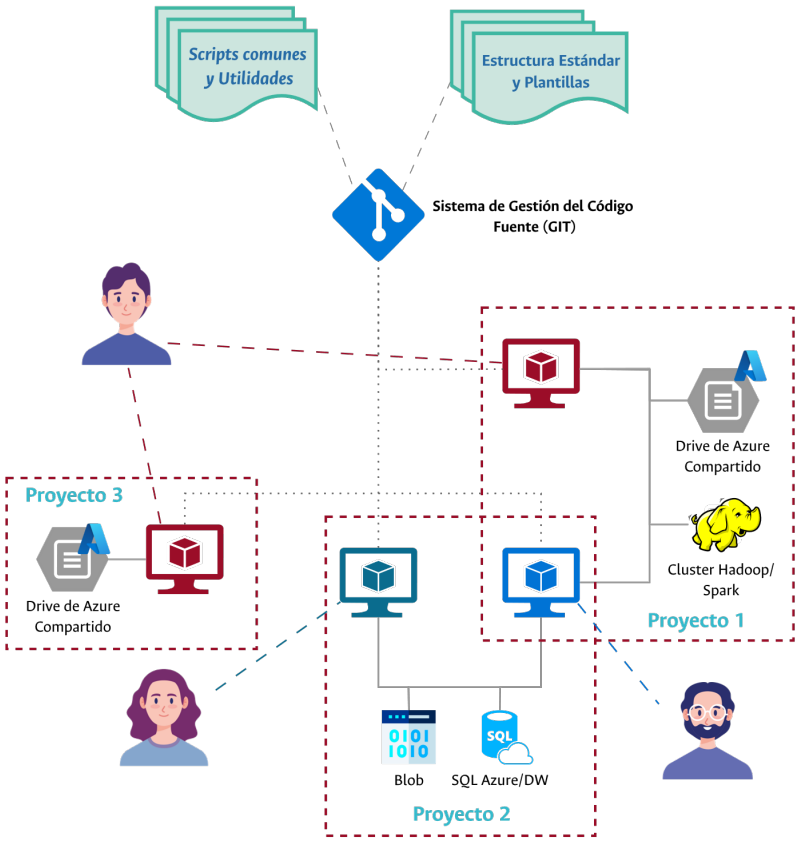


1. Busca definir una estructura (árbol) de directorios para que sea sencillo para los integrantes encontrar información dentro del proyecto.

2. Incluye elementos de Version Control System (VCS) como git, mercurial, subversion, entre otros.

3. También define plataformas para seguimiento de tareas como Jira, Rally, Azure DevOps, entre otros.

Infraestructura y Recursos

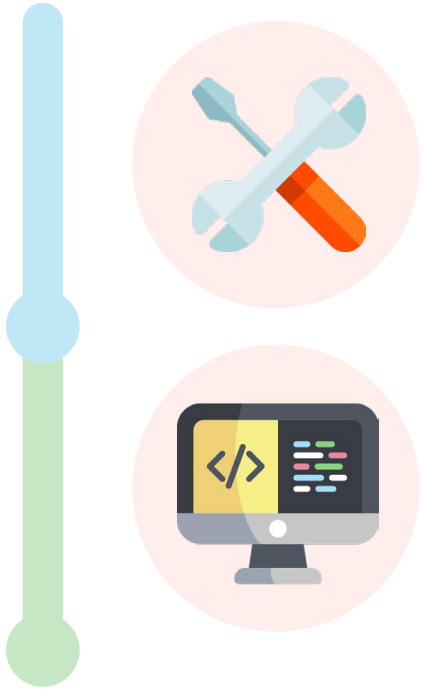


Ejemplo de infraestructura

Permite definir soluciones de infraestructura como

- Sistemas de archivos para almacenar conjuntos de datos.
- Clusters para big data (Spark, Dask, entre otros).
- Bases de datos.
- Servicios de machine learning.

Herramientas y Utilidades



Este componente busca proveer un conjunto inicial de herramientas y utilidades para comenzar a adoptar TDSP en un equipo.

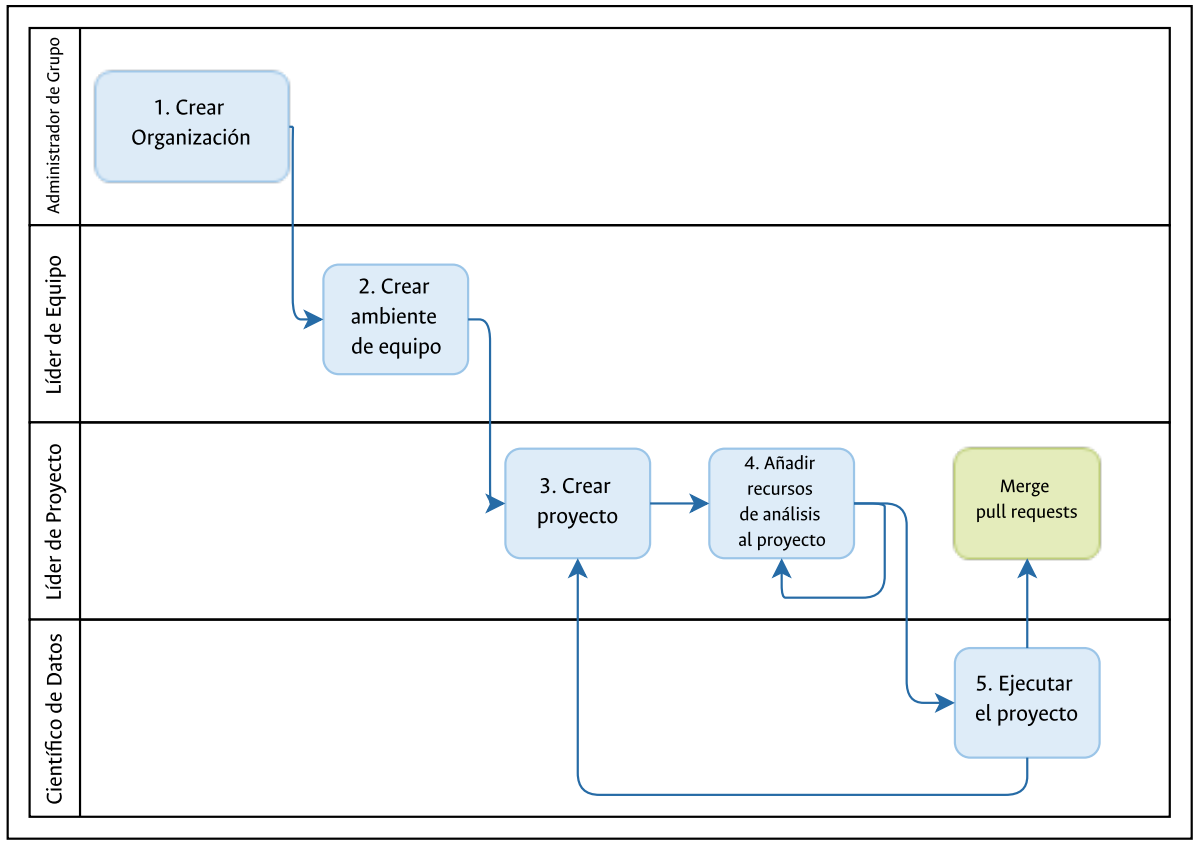
El propósito es definir qué lenguajes de programación, frameworks y herramientas (AutoML, APIs).

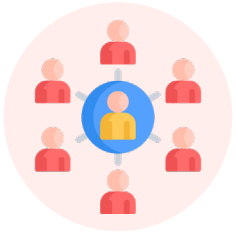
Ejemplos de herramientas

Roles por Unidad

De forma general, TDSP define los siguientes roles:

Roles generales





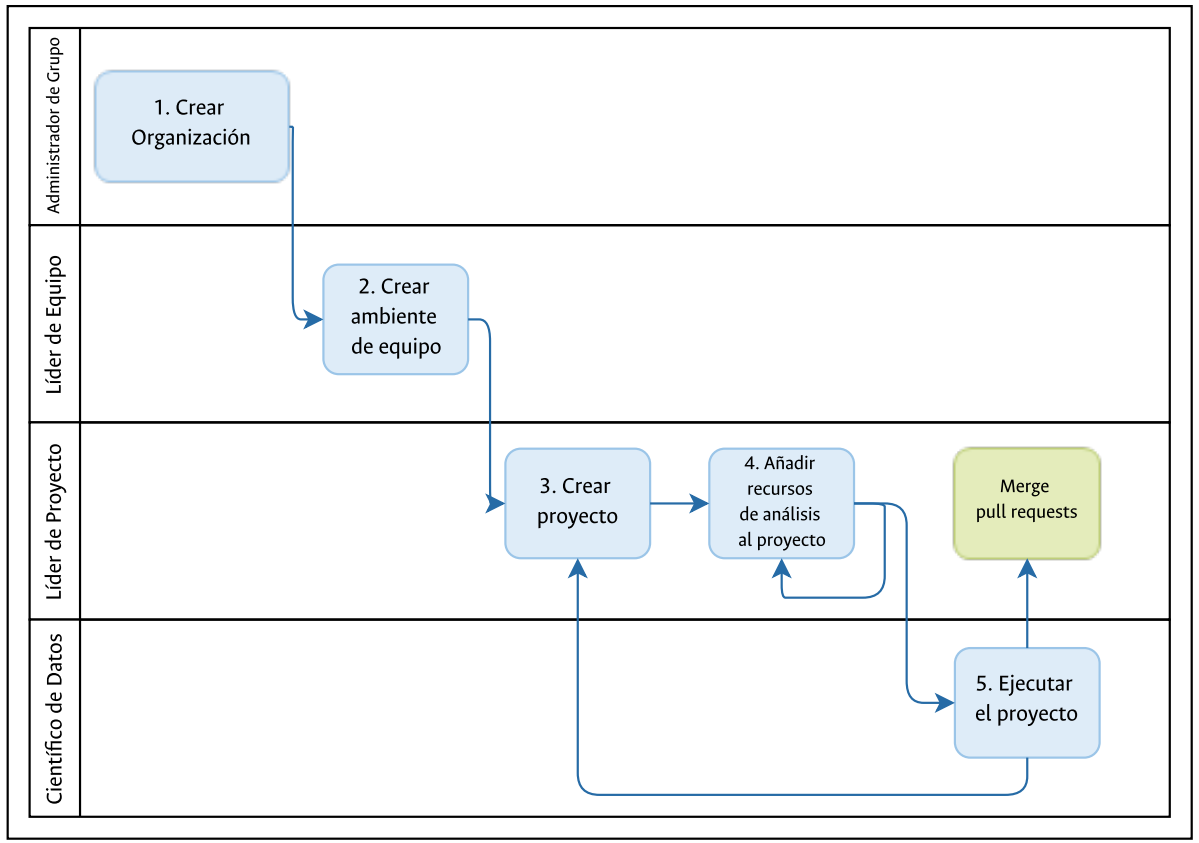
Administrador de grupo: administra toda la unidad de ciencia de datos en una empresa. Una unidad de ciencia de datos puede tener varios equipos, cada uno de los cuales trabaja en varios proyectos de ciencia de datos en distintas verticales comerciales.

Líder de equipo: administra un equipo en la unidad de ciencia de datos de una empresa. Un equipo consta de varios científicos de datos.

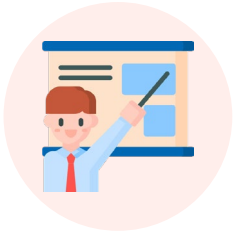
Roles por Unidad

De forma general, TDSP define los siguientes roles:

Roles generales

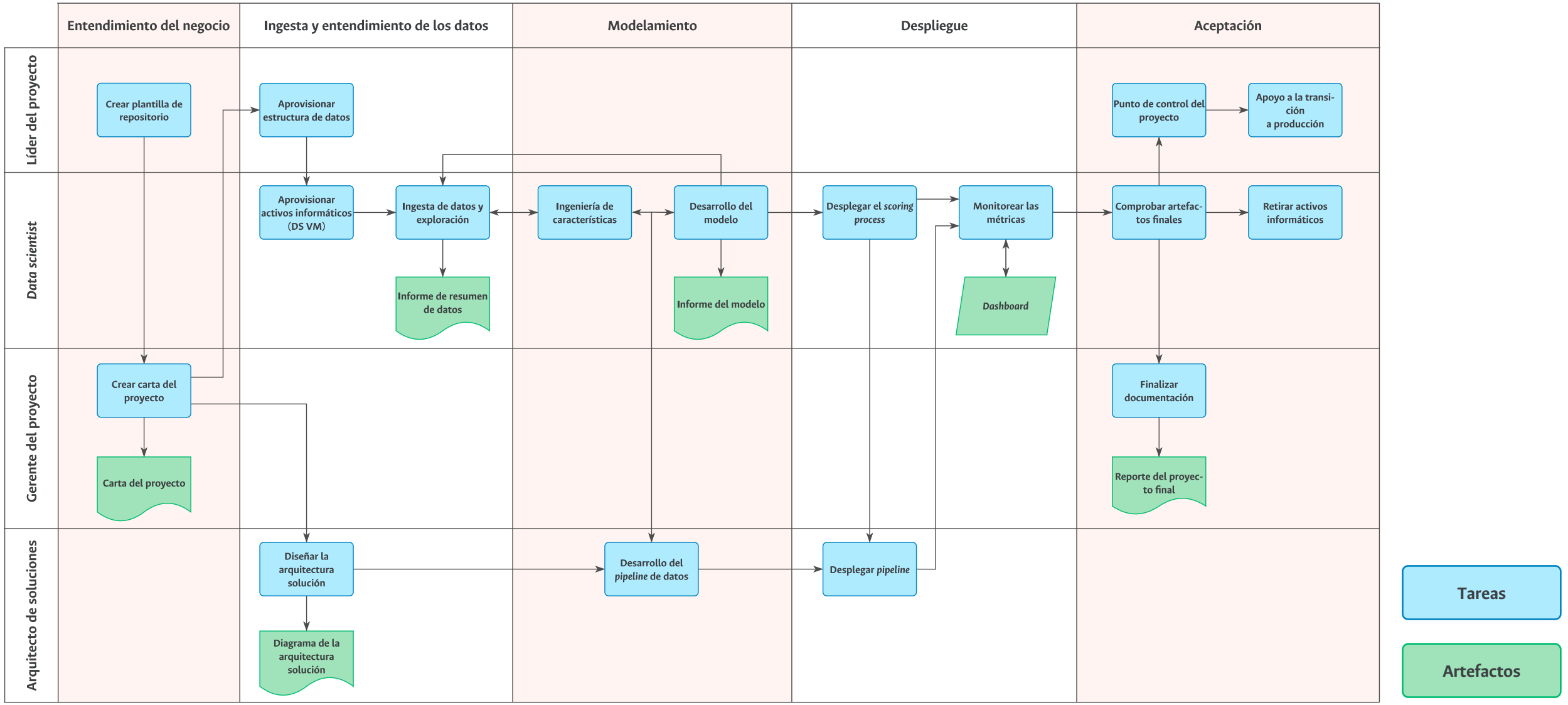


Líder de proyecto: gestiona las actividades diarias de los científicos de datos individuales en un proyecto de ciencia de datos específico.



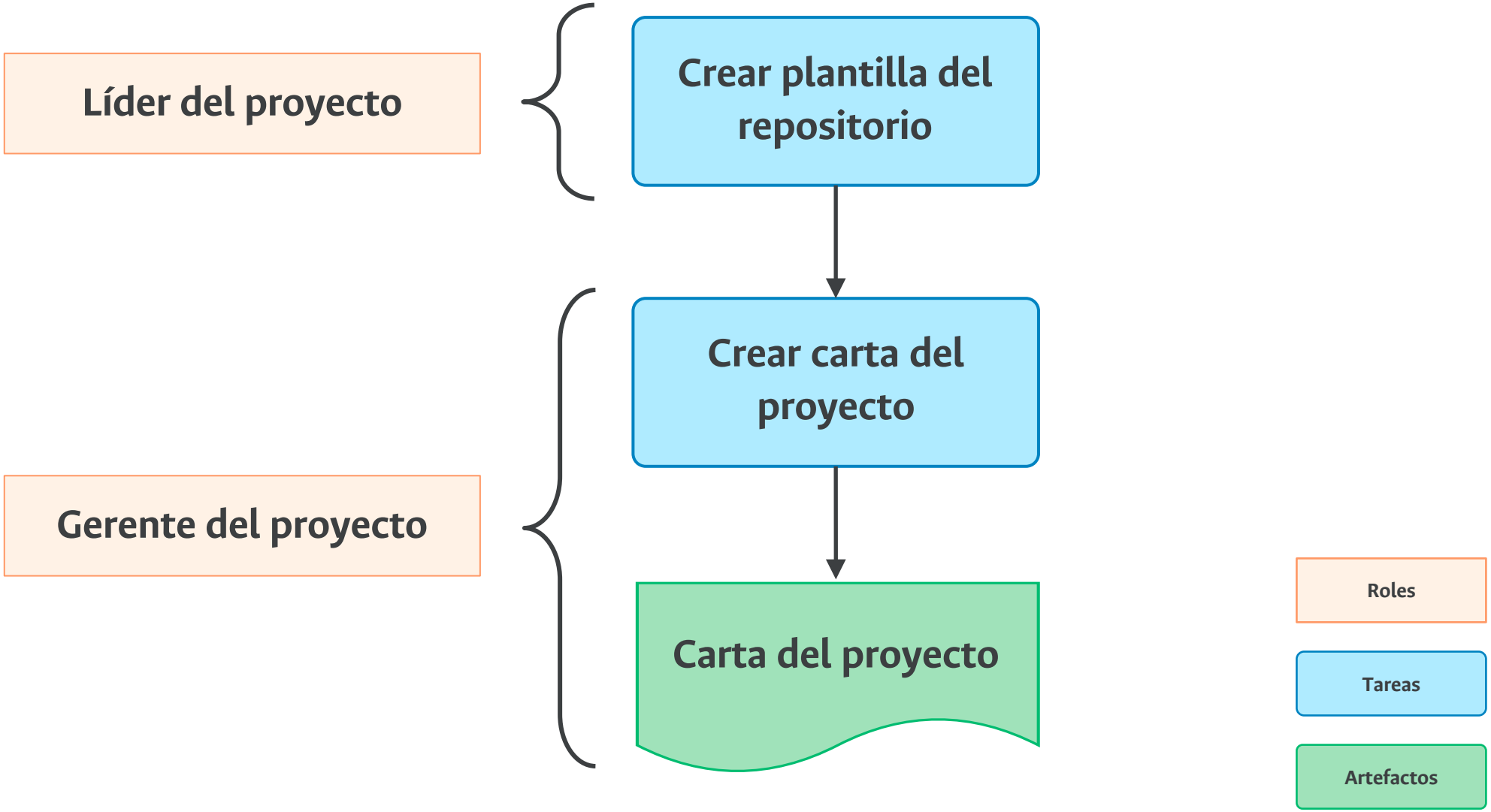
Desarrolladores individuales del proyecto: científicos de datos, analistas de negocios, ingenieros de datos, arquitectos y otros que ejecutan un proyecto de ciencia de datos.

Roles por Equipo



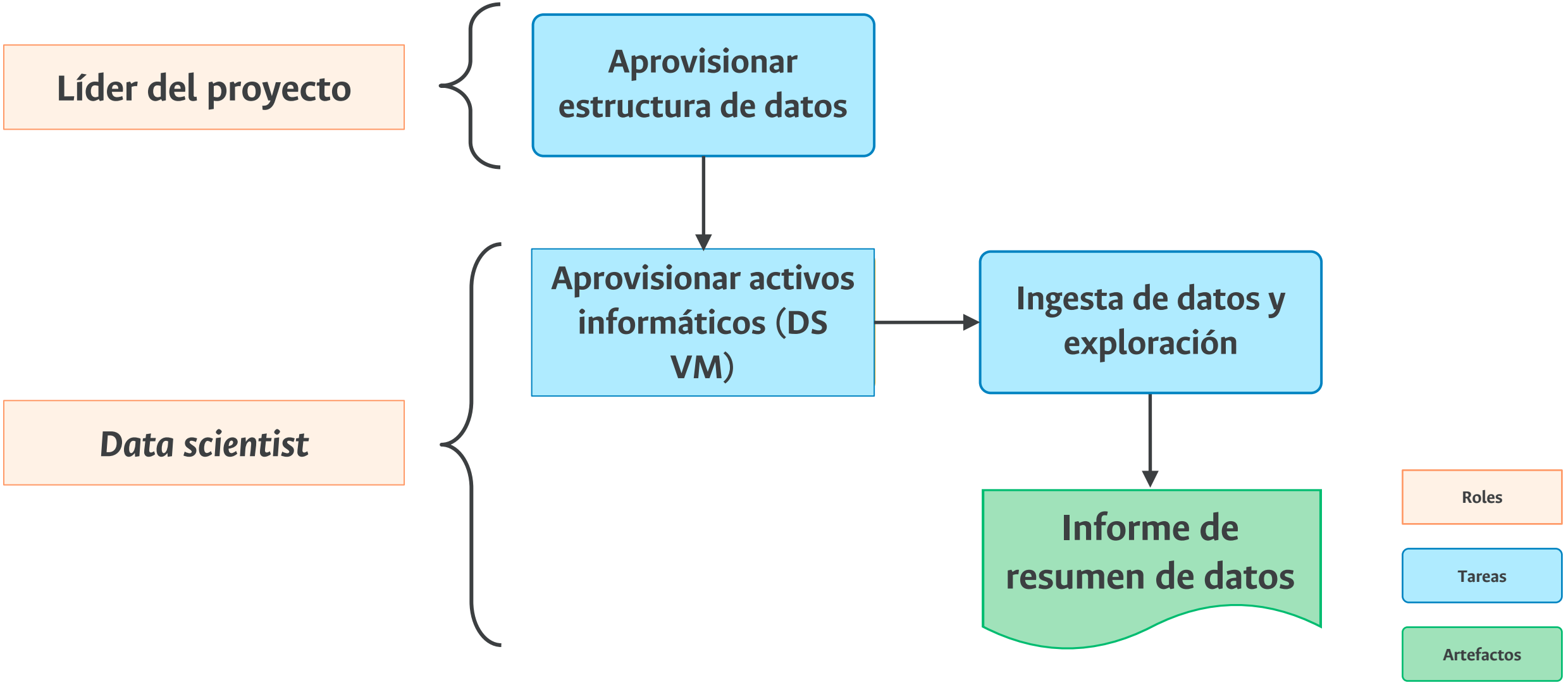
Roles por equipo

Entendimiento del negocio



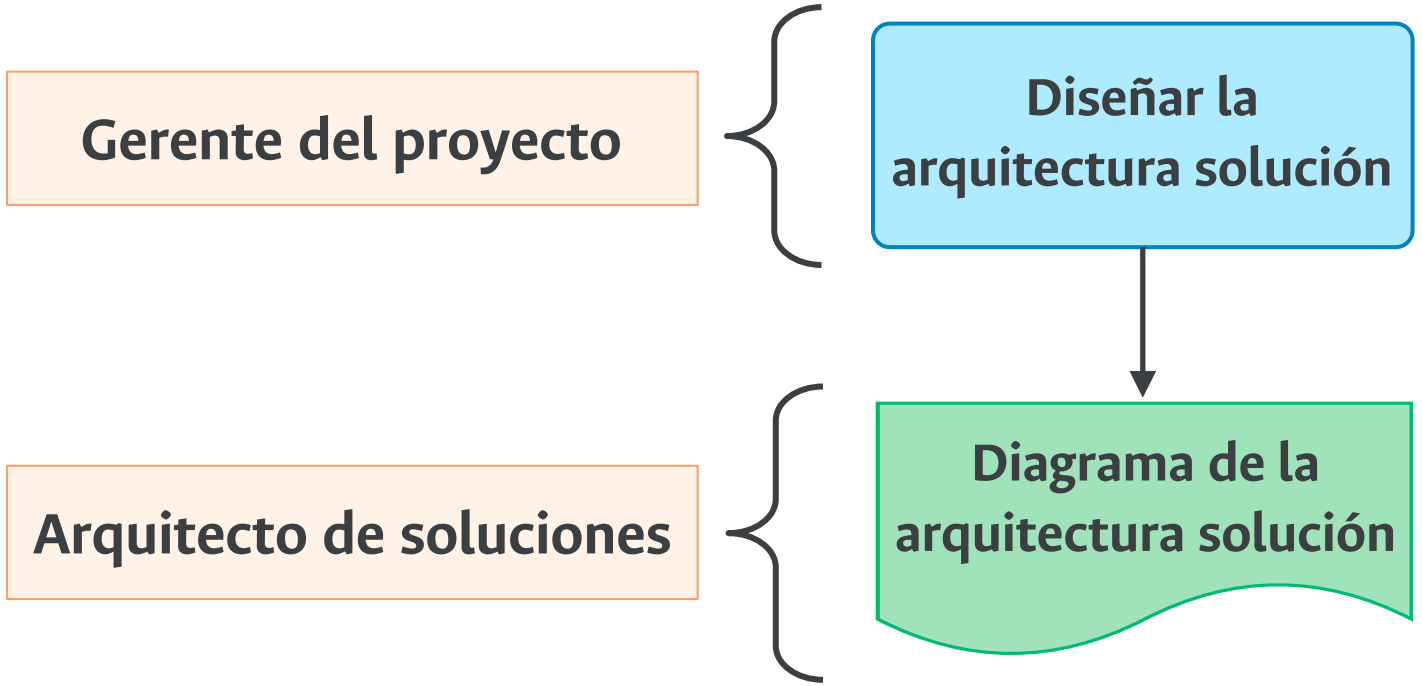
TDSP – Roles y Artefactos

Ingesta y entendimiento de los datos



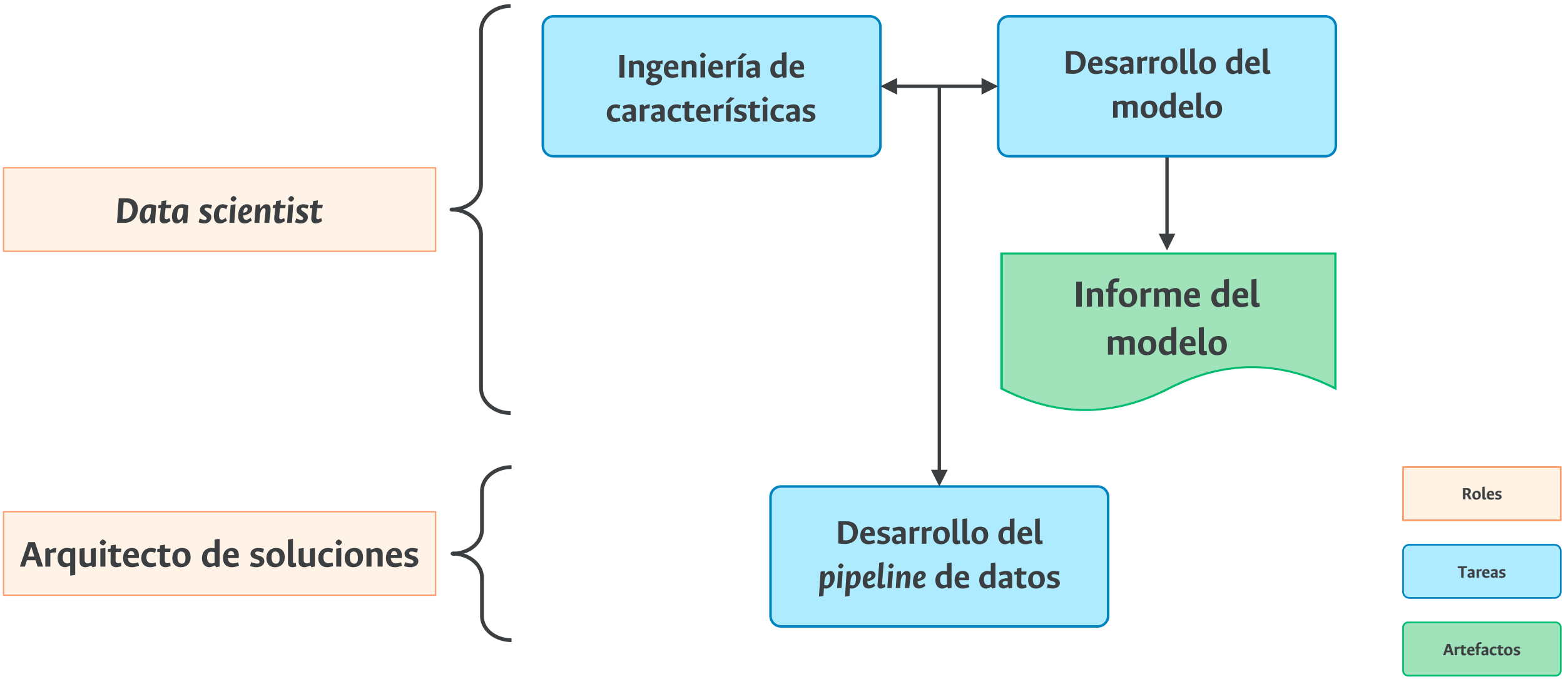
TDSP – Roles y Artefactos

Ingesta y entendimiento de los datos



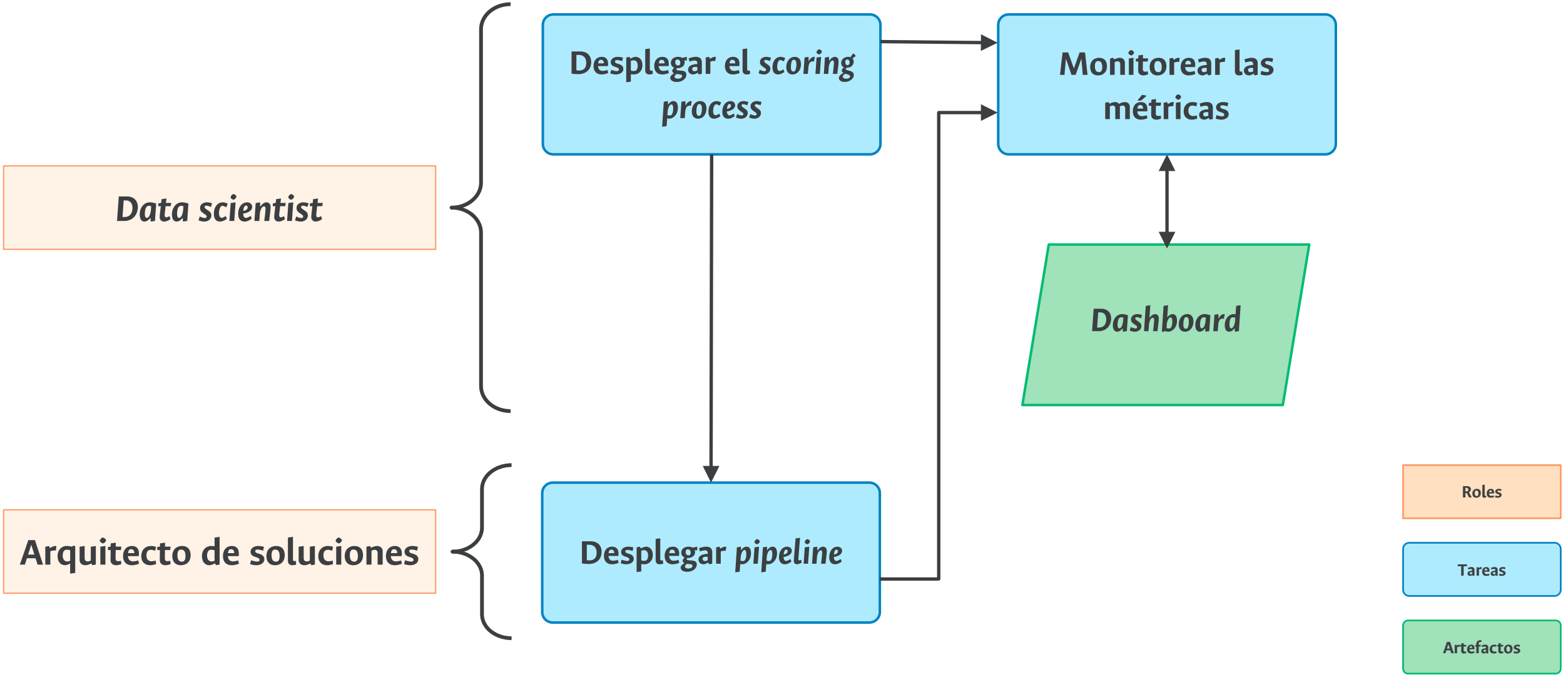
TDSP – Roles y Artefactos

Modelamiento



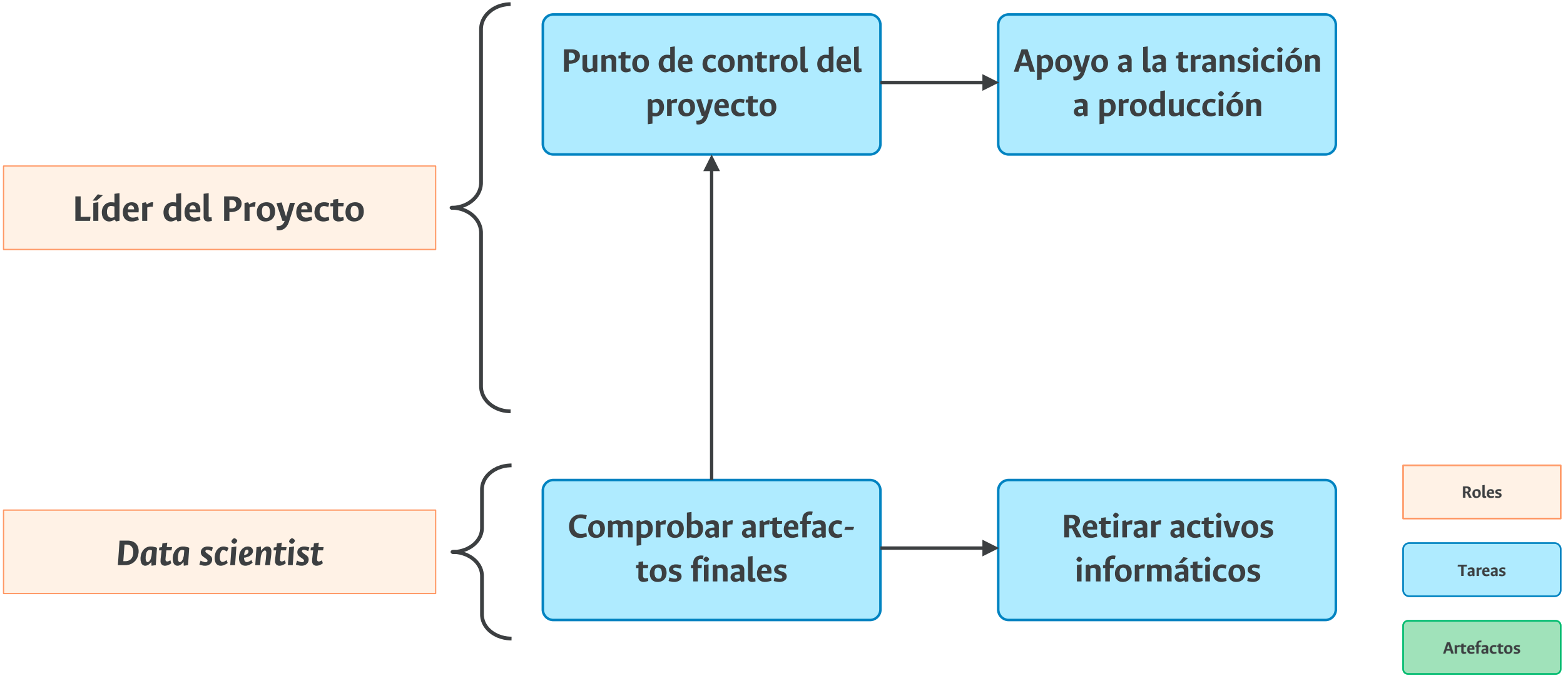
TDSP – Roles y Artefactos

Despliegue



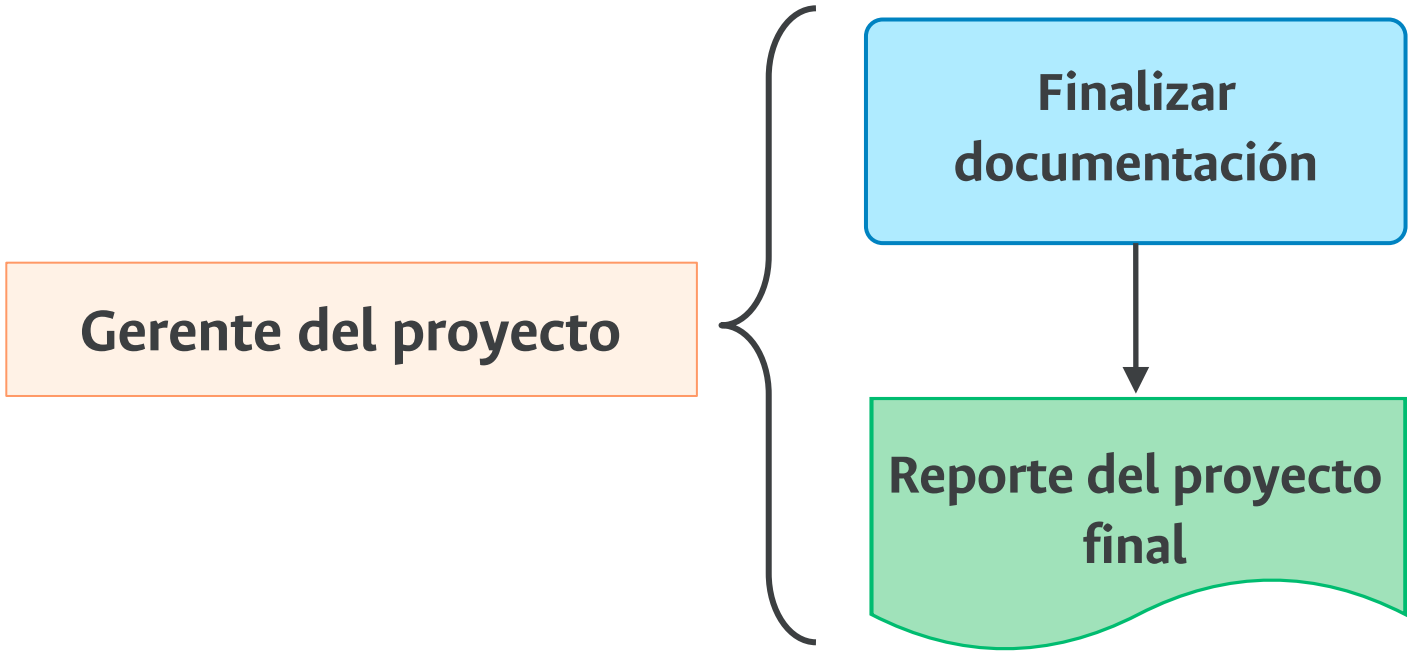
TDSP – Roles y Artefactos

Aceptación



TDSP – Roles y Artefactos

Aceptación

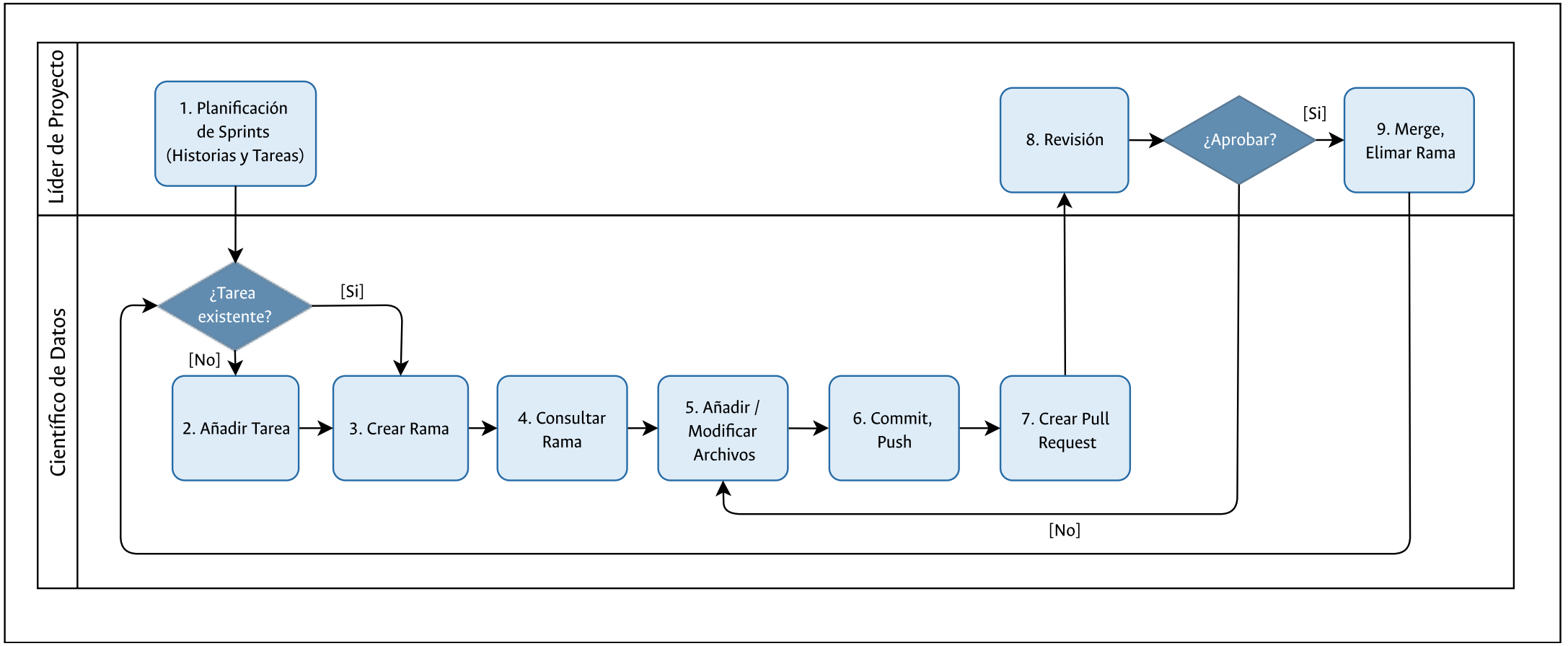


Roles

Tareas

Artefactos

Flujo de Trabajo



Equipo de Ciencia de Datos

Ventajas y Desventajas

Ventajas

- **Ágil**: enfatiza la necesidad de artefactos incrementales.
- **Familiar**: la acumulación de productos, las características, las historias de usuario, los errores, el control de versiones de Git y la planificación de sprints son familiares para quienes están acostumbrados a las prácticas de software comunes.
- **Data Science Native**: TDSP reconoce que la ciencia de datos y la ingeniería de software son diferentes y está diseñado para equipos de ciencia de datos que trabajan en proyectos de producción.
- **Flexible**: TDSP se puede implementar tal como se define o junto con otros enfoques como CRISP-DM.
- **Completo**: debido a su rico enfoque en equipo y documentación detallada, TDSP es posiblemente el enfoque de gestión de proyectos derivado de CRISP más maduro.

Desventajas

- **Sprints fijos**: TDSP aprovecha los sprints de planificación de duración fija con los que luchan muchos científicos de datos.
- **Algunas inconsistencias**: no toda la documentación de Microsoft es coherente.
- **Formación**: requiere conocimientos de desarrollo de software y de ciencia de datos, por lo cual requiere roles más diversos en comparación con las metodologías tradicionales.

Despedida

¡Gracias por su atención!

Jorge Eliécer Camargo Mendoza, PhD.

https://dis.unal.edu.co/~jecamargom/

jecamargom@unal.edu.co

Departamento de Ingeniería de Sistemas e Industrial
Facultad de Ingeniería
Universidad Nacional de Colombia
Sede Bogotá

UNIVERSIDAD NACIONAL DE COLOMBIA

Referencias

- Git. (s. f.). https://git-scm.com/
- Get Started: Data and Model Access. (s. f.). Data Version Control · DVC. https://dvc.org/doc/start/data-management/data-and-model-access
- Get Started: Data Versioning. (s. f.). Data Version Control · DVC. https://dvc.org/doc/start/data-management/data-versioning
- MLflow Tracking – MLflow 2.2.2 documentation. (s. f.-b). https://mlflow.org/docs/latest/tracking.html
- MLflow Models – MLflow 2.2.2 documentation. (s. f.-b). https://mlflow.org/docs/latest/models.html
- MLflow Projects – MLflow 2.2.2 documentation. (s. f.). https://mlflow.org/docs/latest/projects.html
- MLflow Documentation – MLflow 2.2.2 documentation. (s. f.). https://mlflow.org/docs/latest/index.html

Derechos de imágenes

- Flaticon. (s.f.). Idea icon. [Icono]. https://www.flaticon.com/free-icon/idea_2896428
- Flaticon. (s.f.). Binary code icon. [Icono]. https://www.flaticon.com/free-icon/binary-code_2742010
- Flaticon. (s.f.). Predictive models icon. [Icono]. https://www.flaticon.com/free-icon/predictive-models_2103652
- Flaticon. (s.f.). Data transformation icon. [Icono]. https://www.flaticon.com/free-icon/data-transformation_7440397
- Flaticon. (s.f.). Evaluate icon. [Icono]. https://www.flaticon.com/free-icon/evaluate_8089781
- Flaticon. (s.f.). Startup icon. [Icono]. https://www.flaticon.com/free-icon/startup_9119213
- Flaticon. (s.f.). Repeat icon. [Icono]. https://www.flaticon.com/free-icon/repeat_7744716
- Flaticon. (s.f.). Idea icon. [Icono]. https://www.flaticon.com/free-icon/idea_610319
- Flaticon. (s.f.). Data analysis icon. [Icono]. https://www.flaticon.com/free-icon/data-analysis_6260220
- Flaticon. (s.f.). Web development icon. [Icono]. https://www.flaticon.com/free-icon/web-development_5763745
- Flaticon. (s.f.). Idea icon. [Icono]. https://www.flaticon.com/free-icon/idea_8920544
- Flaticon. (s.f.). Problem solving icon. [Icono]. https://www.flaticon.com/free-icon/problem-solving_4133589
- Flaticon. (s.f.). Predictive icon. [Icono]. https://www.flaticon.com/free-icon/predictive_9422873
- Flaticon. (s.f.). Data science icon. [Icono]. https://www.flaticon.com/free-icon/data-science_2103579
- Flaticon. (s.f.). Scrum icon. [Icono]. https://www.flaticon.com/free-icon/scrum_8759187
- Flaticon. (s.f.). Database management icon. [Icono]. https://www.flaticon.com/free-icon/database-management_9672242

Derechos de imágenes

- Flaticon. (s.f.). Refresh icon. [Icono]. https://www.flaticon.com/free-icon/refresh_7838482
- Flaticon. (s.f.). Sitemap icon. [Icono]. https://www.flaticon.com/free-icon/sitemap_3093704
- Flaticon. (s.f.). Automated process icon. [Icono]. https://www.flaticon.com/free-icon/automated-process_2581947
- Flaticon. (s.f.). Customer support icon. [Icono]. https://www.flaticon.com/free-icon/customer-support_1086581
- Flaticon. (s.f.). Idea icon. [Icono]. https://www.flaticon.com/free-icon/idea_1000352
- Flaticon. (s.f.). Database icon. [Icono]. https://www.flaticon.com/free-icon/database_2656281
- Flaticon. (s.f.). Lab Tool icon. [Icono]. https://www.flaticon.com/free-icon/lab-tool_3377205

Créditos

Facultad de
INGENIERÍA

Profesor
Jorge Eliécer Camargo Mendoza, PhD

Asistente docente
Juan Sebastián Lara Ramírez

Coordinador de virtualización
Edder Hernández Forero

Diagramadores PPT
Mario Andrés Rodríguez Triana
Rosa Alejandra Superlano Esquibel

Diseño gráfico
Clara Valeria Suárez Caballero
Milton R. Pachón Pinzón

2025