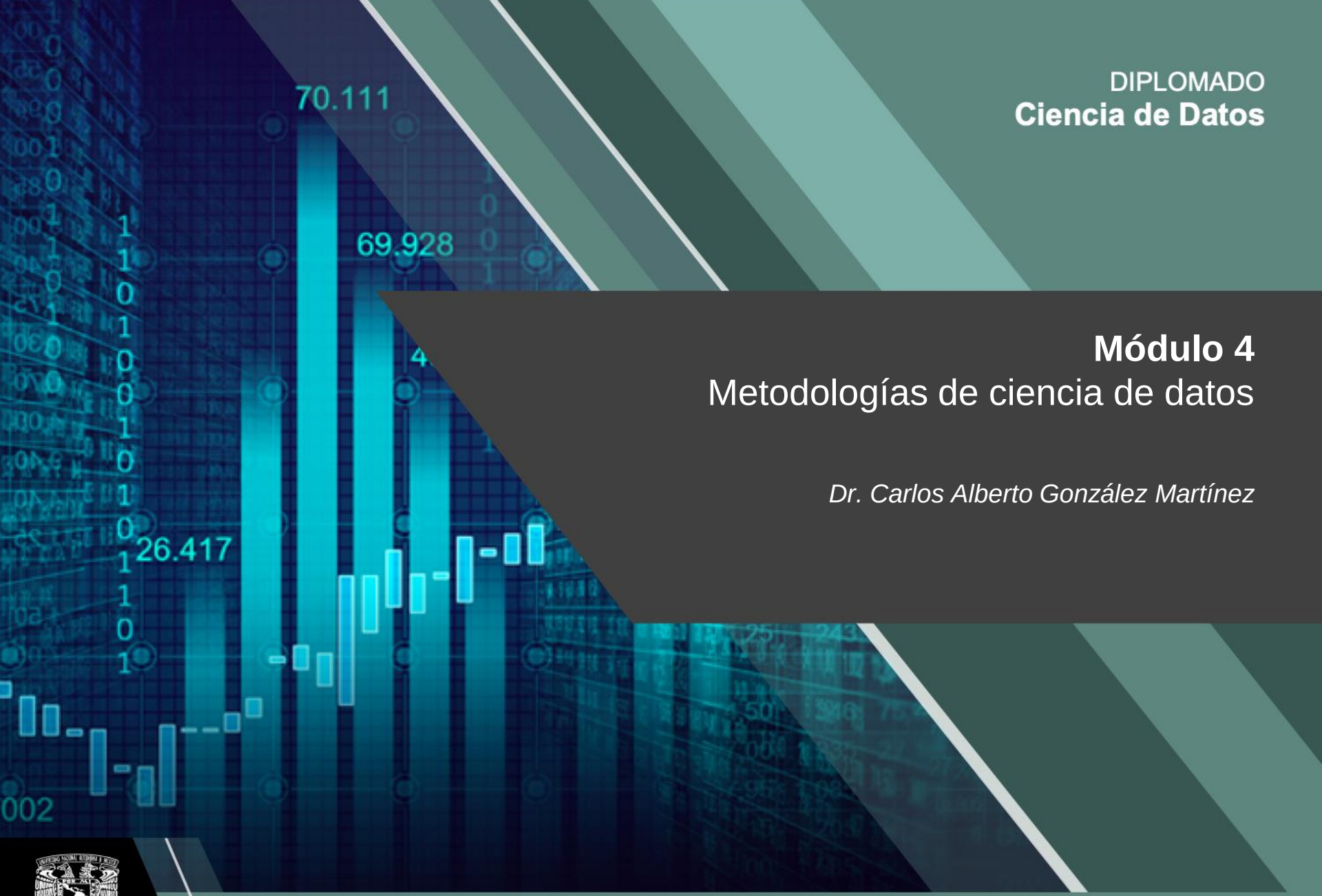


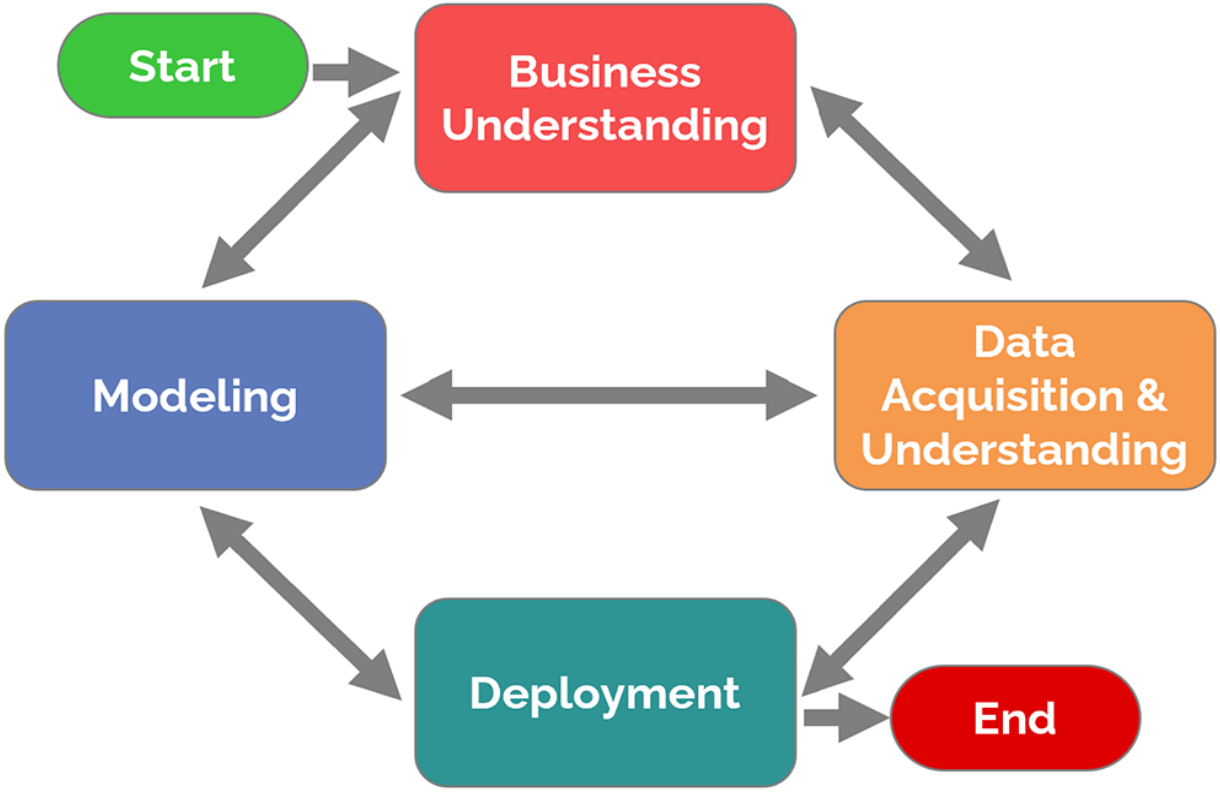
DIPLOMADO
Ciencia de Datos

Módulo 4
Metodologías de ciencia de datos

Dr. Carlos Alberto González Martínez

DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

TEAM DATA SCIENCE PROCESS (TDSP)



gmc

Objetivo

El participante identificará el Team Data Science Process (TDSP), como una metodología de ciencia de datos ágil e iterativa para proporcionar soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente.

Proceso de ciencia de datos en equipo (TDSP)

Contenido

1. Proceso de ciencia de datos en equipo TDSP
2. Principales componentes del TDSP
3. Estructura de proyecto estandarizado
4. Roles y tareas
5. Infraestructura y recursos
6. Herramientas y utilidades

Proceso de ciencia de datos en equipo (TDSP)

1. Proceso de ciencia de datos en equipo TDSP

El proceso de ciencia de datos en equipo (TDSP) proporciona un ciclo de vida para estructurar el desarrollo de los proyectos de ciencia de datos

El ciclo de vida describe las fases principales por las que pasan normalmente los proyectos, a menudo de forma iterativa:

1. Conocimiento del negocio
2. Adquisición y comprensión de los datos
3. Modelado
4. Implementación
5. Aceptación del cliente

Proceso de ciencia de datos en equipo (TDSP)

2. Principales componentes del TDSP

Ciclo de vida de ciencia de datos

El proceso de ciencia de datos en equipo (TDSP) proporciona un ciclo de vida para estructurar el desarrollo de los proyectos de ciencia de datos. En el ciclo de vida se describen todos los pasos que siguen los proyectos correctos.

Aunque esté usando otro ciclo de vida de ciencia de datos, como CRISP-DM, KDD o el proceso personalizado de su organización, puede usar también el TDSP basado en tareas, en el contexto de esos ciclos de vida de desarrollo. En un nivel alto, estas distintas metodologías tienen mucho en común.

Proceso de ciencia de datos en equipo (TDSP)

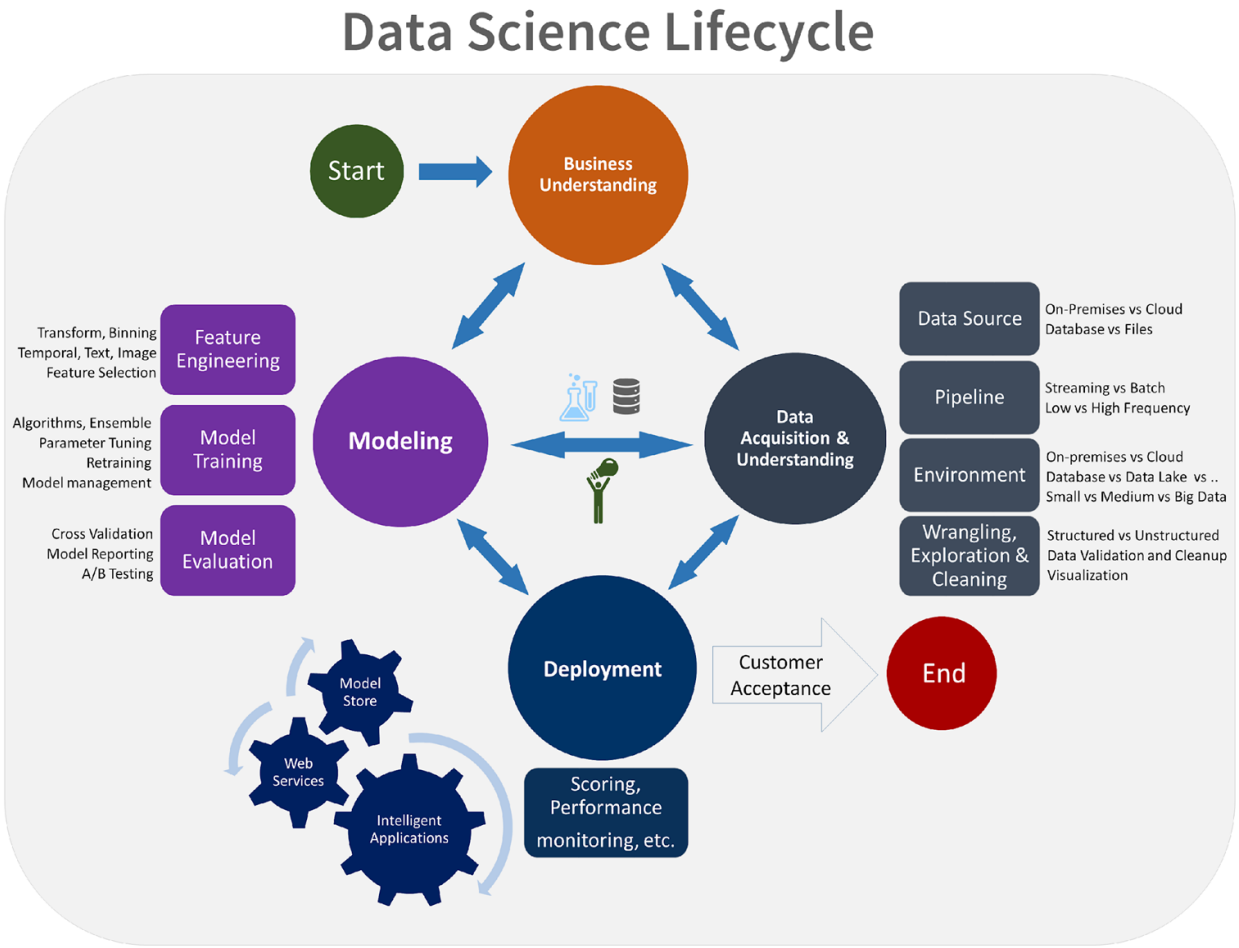
2. Principales componentes del TDSP

El ciclo de vida describe las fases principales por las que pasan normalmente los proyectos, a menudo de forma iterativa:

- Conocimiento del negocio
- Adquisición y comprensión de los datos
- Modelado
- Implementación

La siguiente imagen es una representación visual del **ciclo de vida del proceso de ciencia de datos en equipo**.

Proceso de ciencia de datos en equipo (TDSP)



Proceso de ciencia de datos en equipo (TDSP)

3. Una estructura de proyecto estandarizado

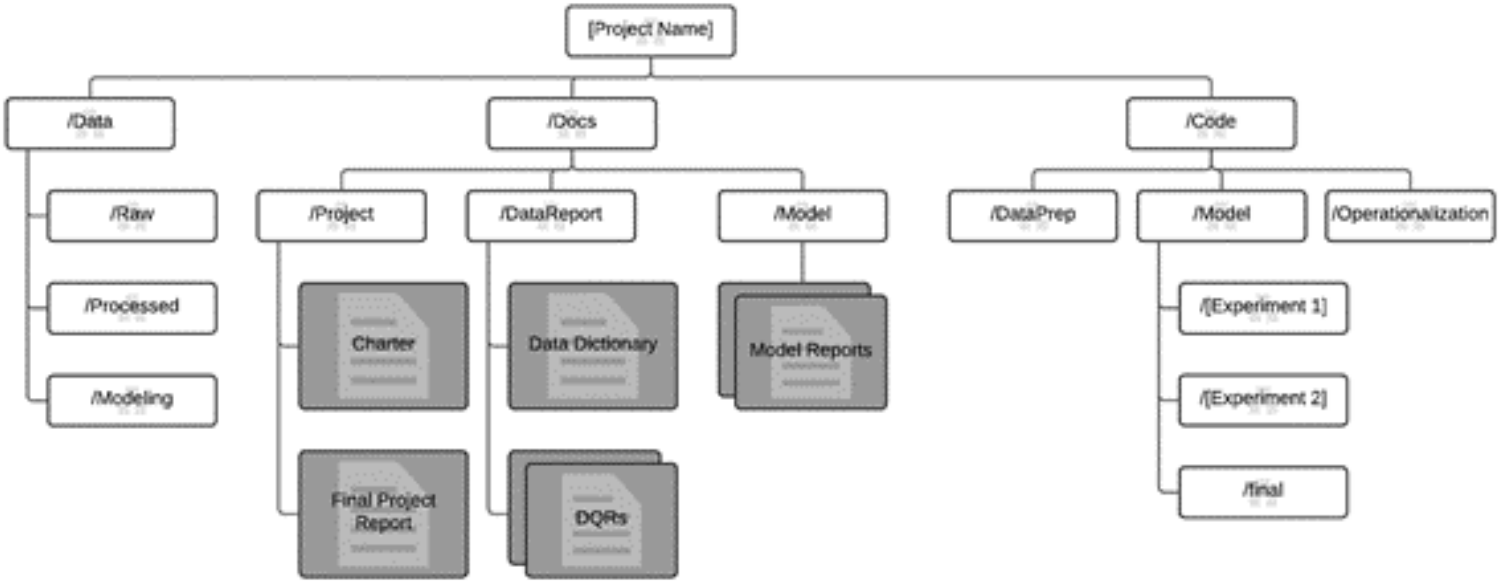
Estructura de proyecto estandarizado. Cuando todos los proyectos comparten una estructura de directorio y usan plantillas para los documentos de proyecto, resulta fácil para los miembros del equipo encontrar información sobre sus proyectos.

Todo el código y los documentos se almacenan en un sistema de control de versiones (VCS), como Git, TFS o Subversion, para permitir la colaboración en equipo. El seguimiento de las tareas y las características en un sistema de seguimiento de proyectos ágil, como Jira, Rally y Azure DevOps, permite seguir más de cerca el código para conocer sus características individuales.

Proceso de ciencia de datos en equipo (TDSP)

3. Una estructura de proyecto estandarizado

Estructura típica de directorio usada en GitHub.



Proceso de ciencia de datos en equipo (TDSP)

4. Roles y tareas

Definición y tareas de los cuatro roles de TDSP

1. Administrador de grupo: administra la unidad de ciencia de datos completa en una empresa. Una unidad de ciencia de datos podría tener varios equipos, cada uno de ellos trabajando en varios proyectos de ciencia de datos en segmentos verticales de negocio distintos. Un Administrador de grupo puede delegar sus tareas en un suplente, pero no cambian las tareas asociadas al rol.

2. Responsable de equipo: administra un equipo de la unidad de ciencia de datos de una empresa. Un equipo está formado por varios científicos de datos. En una unidad de ciencia de datos reducida, el administrador de grupo y el responsable de equipo podrían ser la misma persona.

Proceso de ciencia de datos en equipo (TDSP)

4. Roles y tareas

3. Responsable de proyecto: administra las actividades diarias de los científicos de datos en un proyecto de ciencia de datos específico.

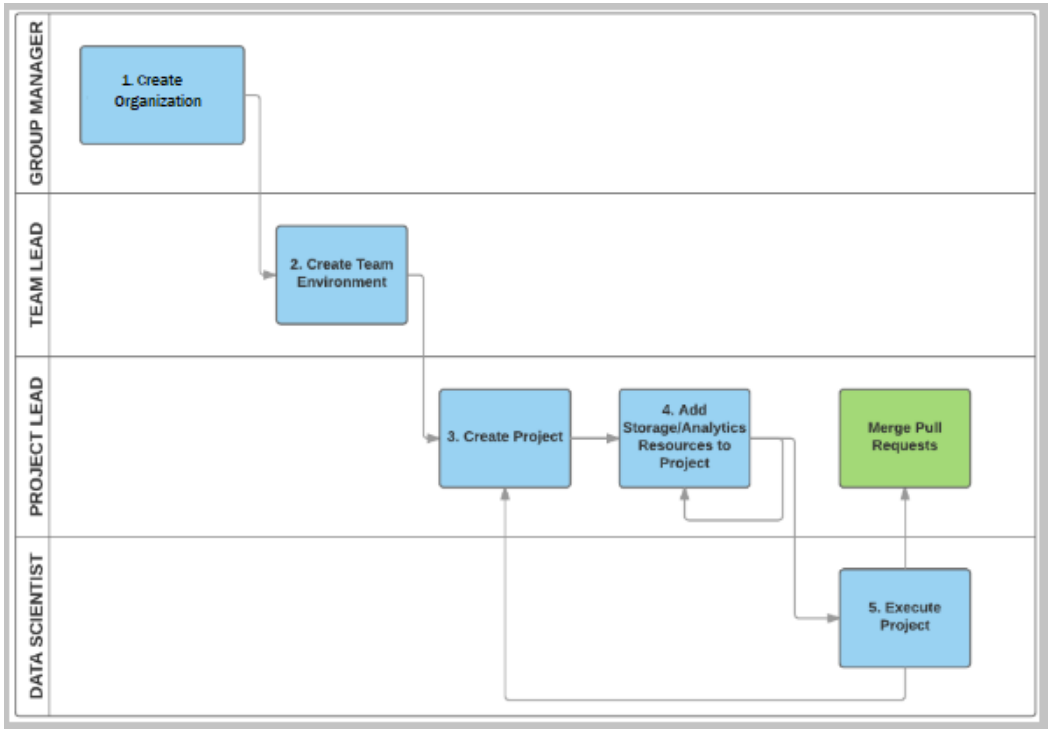
4. Colaboradores individuales del proyecto: científicos de datos, analistas de negocios, ingenieros de datos, arquitectos y otros colaboradores que ejecutan un proyecto de ciencia de datos.

Nota: En función de la estructura y el tamaño de una empresa, una sola persona puede desempeñar más de un rol o un grupo de personas podría ocupar un solo rol.

Proceso de ciencia de datos en equipo (TDSP)

4. Roles y tareas

En el diagrama siguiente se muestran las tareas de nivel superior para cada rol del proceso de ciencia de datos en equipo.



gmc

Proceso de ciencia de datos en equipo (TDSP)

4. Roles y tareas

Flujo de trabajo de ejecución del proyecto de ciencia de datos

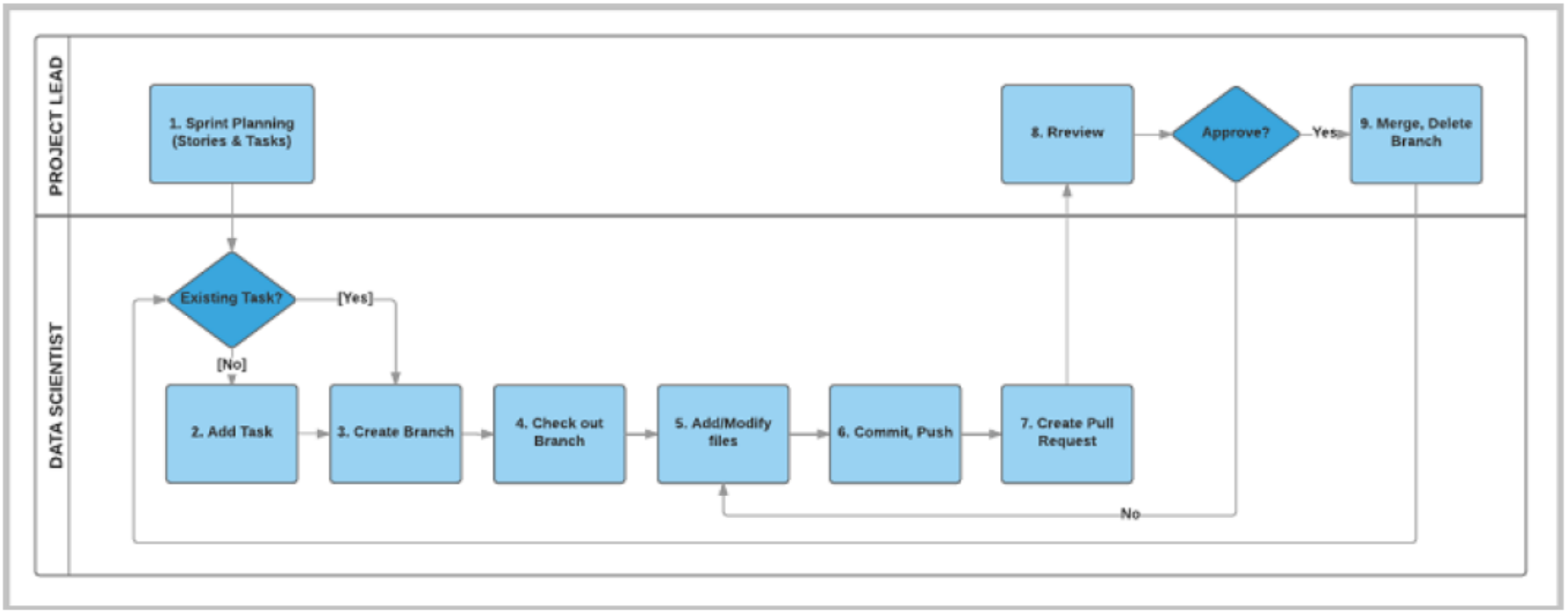
Los responsables de proyecto y los responsables de equipo, pueden crear elementos de trabajo para realizar el seguimiento de todas las tareas y las fases de un proyecto, de principio a fin.

La siguiente ilustración muestra el flujo de trabajo de TDSP para la ejecución del proyecto:

Proceso de ciencia de datos en equipo (TDSP)

4. Roles y tareas

Flujo de trabajo de ejecución del proyecto de ciencia de datos



gmc

Proceso de ciencia de datos en equipo (TDSP)

4. Roles y tareas

Flujo de trabajo de ejecución del proyecto de ciencia de datos

Los pasos del flujo de trabajo se pueden agrupar en tres actividades:

- Los responsables de proyecto realizan el planeamiento de sprints.
- Los científicos de datos desarrollan artefactos en ramas de git para tratar los elementos de trabajo.
- Los responsables de proyecto u otros miembros del equipo realizan revisiones de código y combinan ramas de trabajo en la rama primaria.

Proceso de ciencia de datos en equipo (TDSP)

5. Infraestructura y recursos recomendados para proyectos de ciencia de datos.

La estructura de directorio se puede utilizar o clonar desde GitHub.
TDSP proporciona recomendaciones para administrar análisis compartido e infraestructura de almacenamiento,
por ejemplo:

- Sistemas de archivos en la nube para almacenar conjuntos de datos
- Databases
- Clústeres de macrodatos (SQL o Spark)
- Servicio de aprendizaje automático

Proceso de ciencia de datos en equipo (TDSP)

5. Infraestructura y recursos recomendados para proyectos de ciencia de datos

La infraestructura de análisis y almacenamiento, donde se almacenan los conjuntos de datos sin procesar y los procesados, puede estar en la nube o en un entorno local. Esta infraestructura permite un análisis reproducible.

También evita la duplicación, lo que puede llevar a incoherencias y costos de infraestructura innecesarios.

Proceso de ciencia de datos en equipo (TDSP)

6. Herramientas y utilidades recomendadas para la ejecución de proyectos

En la mayoría de las organizaciones, la introducción de procesos presenta ciertos desafíos. Las herramientas proporcionadas para implementar el proceso y el ciclo de vida de ciencia de datos, ayudan a reducir las barreras a su adopción y la normalizan. TDSP proporciona un conjunto inicial de herramientas y scripts para impulsar la adopción de TDSP dentro de un equipo.

También ayuda a automatizar algunas de las tareas comunes del ciclo de vida de ciencia de datos, como la exploración de datos y el modelado de línea de base.

Proceso de ciencia de datos en equipo (TDSP)

6. Herramientas y utilidades recomendadas para la ejecución de proyectos

Estos recursos se pueden aprovechar luego en otros proyectos dentro del equipo o en la organización.

Amazon, Google y Microsoft proporcionan herramientas extensas en Machine Learning, que admiten marcos de código abierto (Python, R, ONNX y aprendizaje profundo común) y también herramientas propias.

Referencias bibliográficas

Microsoft, (2020). ¿Qué es el Proceso de ciencia de datos en equipo (TDSP)?. Recuperado de https://opdhsblobprod01.blob.core.windows.net/contents/4a6d75bb3af747de838e6ccc97c5d978/81a99b988dc9d5d0c39194dce869a507?sv=2018-03-28&sr=b&si=ReadPolicy&sig=00Yck17ZfFgjP1B2XWB%2BC5qYYVhLFT%2B7iKK3SV8Y%2B2Y%3D&st=2021-06-18T17%3A31%3A42Z&se=2021-06-19T17%3A41%3A42Z

Contacto

Carlos Alberto González Martínez

Jefe de departamento de correlaciones, cruces y alertas (C5i)

gmcmmiv@hotmail.com