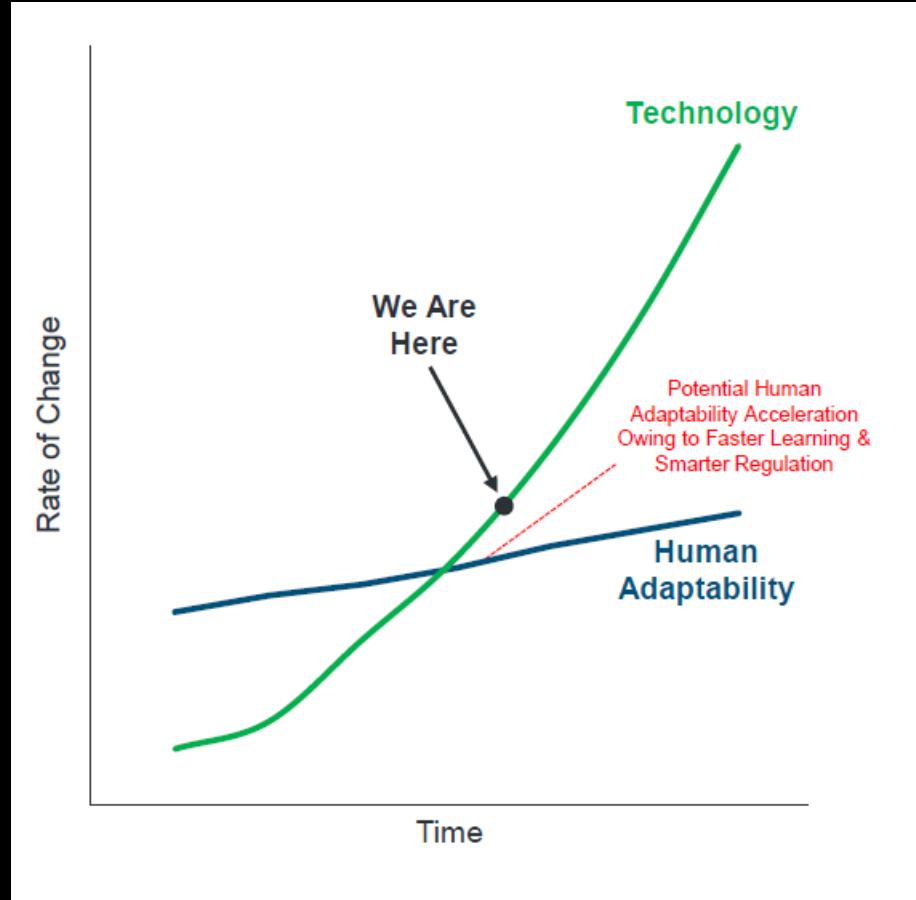


Introducción a Big Data

Conceptos básicos



Cambio Tecnológico > Adaptabilidad Humana



La capacidad de los seres humanos para adaptarse al cambio tecnológico está aumentando, pero no sigue el ritmo de la innovación científica y tecnológica.

Para superar la fricción resultante, los humanos pueden adaptarse mediante el desarrollo de habilidades que permitan un aprendizaje más rápido y una iteración y experimentación más rápidas.

Astro Teller – X, The Moonshot Factory
Adaptado de Thomas Friedman: Gracias por llegar tarde, 2016

La tecnología sigue creciendo y lo seguirá haciendo durante varios años, el protagonista: los datos, llamados el nuevo petróleo. Gran parte del crecimiento de los últimos años podría describirse libremente como "creación de valor a partir de los datos".



El valor podría significar aumentar los ingresos por ventas, reducir los costos evitables, mejorar la satisfacción del paciente, enfocarse en clientes y prospectos de alto valor, crear políticas para el bien social más amplio y mucho, mucho más.



Los datos, el nuevo petróleo del siglo XXI

(Andreas Weigend, ex Chief Scientist de Amazon y experto en Big Data)

Si tu empresa no se puede permitir el tener recursos ociosos ¿por qué podría permitirse el no explotar sus datos?

Pero es así

Un estudio presentado por EMC, compañía de Cloud Computing, Big Data y Soluciones IT, desvela que:



de las empresas reconoce que no sabe cómo convertir todos sus datos en información útil.



está siempre conectado y puede tomar sus decisiones basándose en esa información en tiempo real.



dice que puede extraer conocimiento de los datos.



que reconoce que no utiliza sus datos de manera eficaz o que se encuentran desbordados por una sobrecarga de información.



de organizaciones se consideran "buenas" o "muy buenas" a la hora de transformar los datos en conocimiento e información útil para su negocio.

¿Cómo obtener valor de los datos de tu empresa y del Big Data?

Construye modelos predictivos que te ayuden a ser más eficiente:



Hay una enorme superposición en las áreas donde se crea valor a partir de los datos. Observe la siguiente lista de áreas temáticas superpuestas:

- Business Intelligence (BI)
- Visual BI, Analytics
- Visual Analytics
- Business Analytics
- Data Analytics
- Predictive Analytics
- Prescriptive Analytics
- Advanced Analytics
- Text Analytics
- Graph Analytics
- Social Analytics
- Network Analytics
- Modern Analytics
- Directed Acyclic Graph (DAG) Analytics
- Statistics
- Optimization
- Data Mining
- Data Modeling
- ML
- Big Data
- Data Science
- Decision Science
- (Enterprise or Business) Decision Management
- Business Process Management
- Data Engineering
- AI
- Computational Intelligence
- Auto ML
- Management Science
- Linear and Mathematical Programming
- Deep Learning, Informatics
- Decision Science
- Muchas otras

Explosión de los datos

El volumen de datos recopilados sigue creciendo exponencialmente , un estudio del IDC (International Data Corporation), informó (Reinsel et al., 2017):

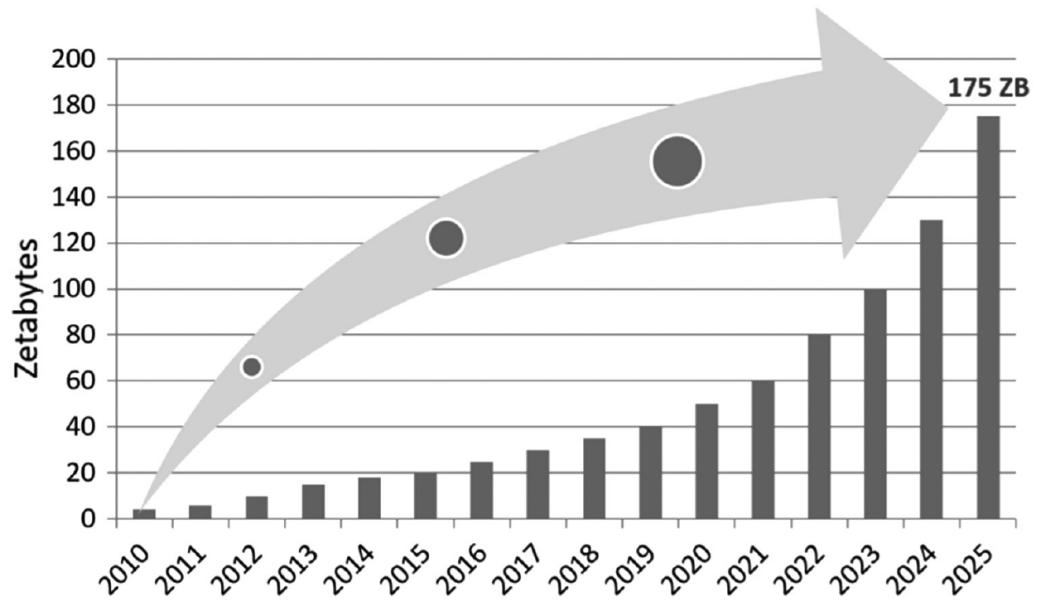
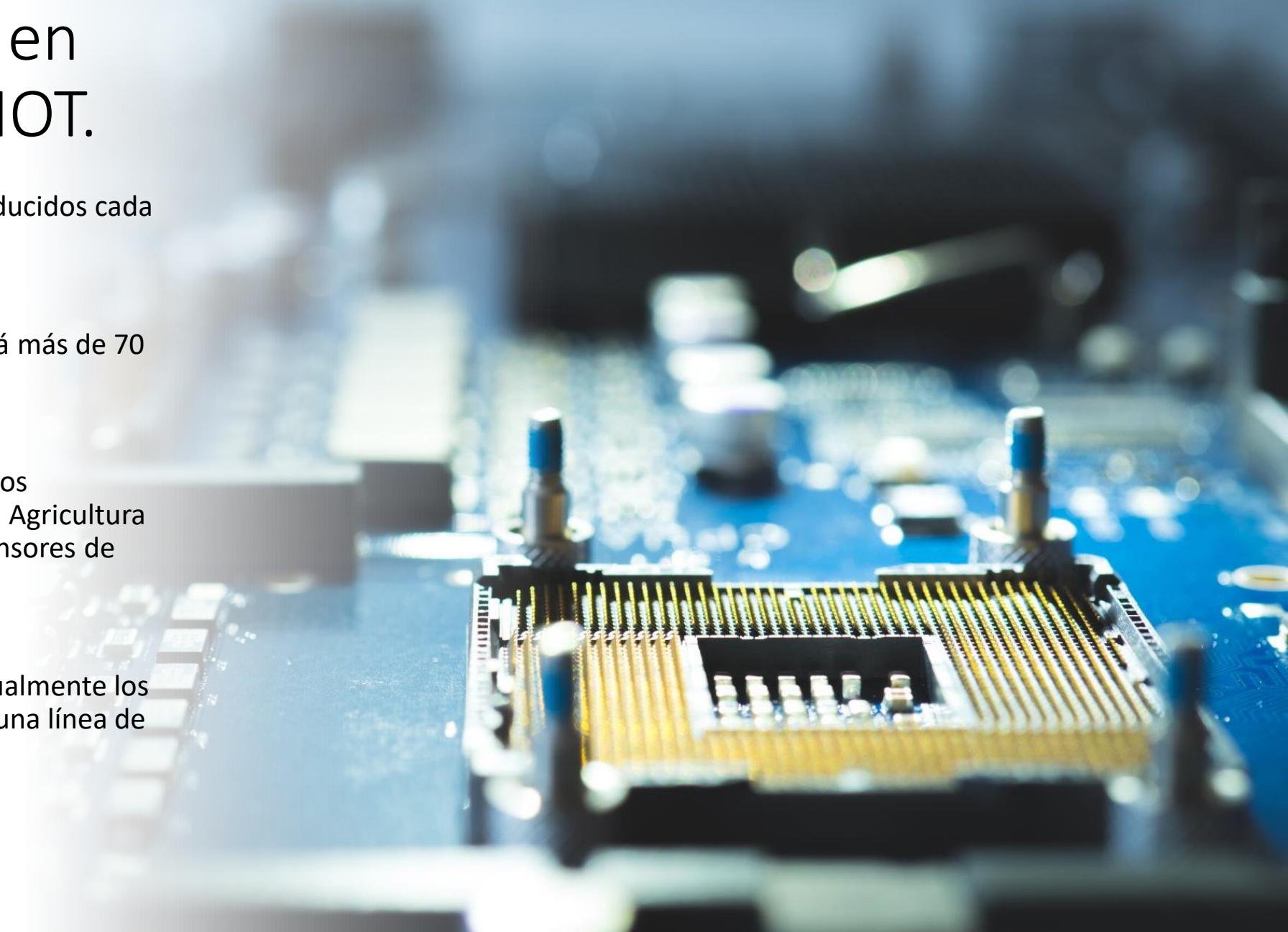


Figure 1.1 Annual size of global data (Adapted from Reinsel et al. 2017, Data Age 2025; The Digitization of the World [IDC Nov 2018]).

- Los datos globales crecerán de 33 Zettabytes en 2018 a 175 Zettabytes para 2025 (un Zettabyte son 1,000,000,000,000,000,000 bytes).
- En 2025 habrá 150 mil millones de dispositivos creando datos en tiempo real. Muchos serán dispositivos de sensores en ciudades inteligentes.

Similarmente CISCO reportó, en el contexto de IOT.

- 5 quintillones de bytes de datos producidos cada día (5×10^{30} bytes/día)
- Para el año 2030, el IoT comprenderá más de 70 mil millones de dispositivos conectados.
- Estos dispositivos incluyen dispositivos inteligentes para el hogar y el automóvil, Agricultura 3.0 que incluye tractores autónomos, sensores de campo, datos satelitales y mucho más.
- Tomaría toda una vida analizar manualmente los datos producidos por un solo sensor en una línea de ensamblaje de fabricación.



La Revista Harvard Bussiness Review también reportó que:

- Menos de la mitad de los datos estructurados se utilizan activamente en la toma de decisiones de negocio
- Menos del 1 % de los datos no estructurados se analizan en absoluto



- La recopilación de datos es costosa. Se necesita dinero para extraer los datos también para la creación de rutas o conductos (pipelines) hacia las bases de datos o lagos de datos a través del proceso de extracción, transformación y carga (ETL).
 - Los costos de almacenamiento son otro factor no solo la compra o el alquiler de la infraestructura, sino la mano de obra especializada asociada con mantener todo en funcionamiento.
 - Según Forbes (Meehan, 2019) se estima que hasta el 90% de los datos nunca se analizan. De hecho, es tan grande que Gartner ha acuñado un término llamado "datos oscuros". Dark Data describe los activos de información que las organizaciones recopilan, procesan y almacenan durante las actividades comerciales regulares, pero que generalmente no se utilizan para otros fines.



Tenemos un problema: Todos estos datos se recopilan a un gran costo, pero no estamos obteniendo valor de ellos.

De allí la necesidad de los roles vinculados al Big Data entre ellos el Ingeniero de Datos , el Científico de Datos y el AI/ML Engineer

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)



Rol más generalista, PHD en especialidades como Astrofísica, Matemáticas, Ciencias Computacionales con habilidades para manipular e interpretar datos

Is Data Scientist Still the Sexiest Job of the 21st Century?

by Thomas H. Davenport and DJ Patil

July 15, 2022



HBR Staff/StudioM1/Moritz Otto/Getty Images

Los científicos de datos tienen ahora distintos roles según el objetivo empresarial y se les piden más habilidades técnicas y blandas.



¿Qué es Ciencia de Datos?

Consiste en la aplicación de varios conocimientos interdisciplinarios y herramientas tecnológicas con la finalidad de convertir los datos en valor de negocio, o convertir los datos en conocimiento de valor que conduzca a la toma de decisiones de negocios.



• ¿Cuál es su finalidad?

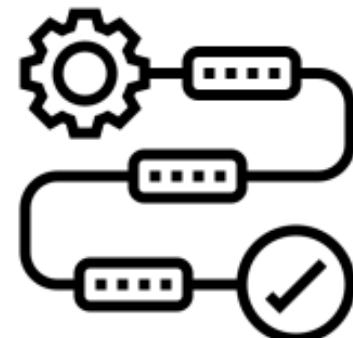
- Tomar decisiones y crear estrategias de negocio.
- Crear productos de software más inteligentes y funcionales.





¿De qué trata este proceso?

- Obtención de los datos.
- Transformar y limpiar los datos.
- Explorar, analizar y visualizar datos.
- Usar modelos de machine learning*.
- Integrar datos e IA a productos de software.



*Inteligencia artificial. No siempre es necesario usarla.

- **Reducción de coste.** Las grandes tecnologías de datos, aportan importantes ventajas en términos de costes cuando se trata de almacenar grandes cantidades de datos, además de identificar maneras más eficientes de hacer negocios.
- **Más rápido, mejor toma de decisiones.** Con la velocidad actual, combinada con la capacidad de analizar nuevas fuentes de datos, las empresas pueden analizar la información inmediatamente y tomar decisiones.
- **Nuevos productos y servicios.** Con la capacidad de medir las necesidades de los clientes y la satisfacción a través de análisis viene el poder de dar a los clientes lo que quieren. Con la analítica de Big Data, más empresas están creando nuevos productos para satisfacer las necesidades de los clientes.



DE DATOS A DASHBOARDS

Tenemos idea de hacer “algo con muchos datos”

Los manipulamos y transformamos para calcular procesos, eficiencias, gastos...

Ejemplo rápido: Ingresos vs Gastos personales.



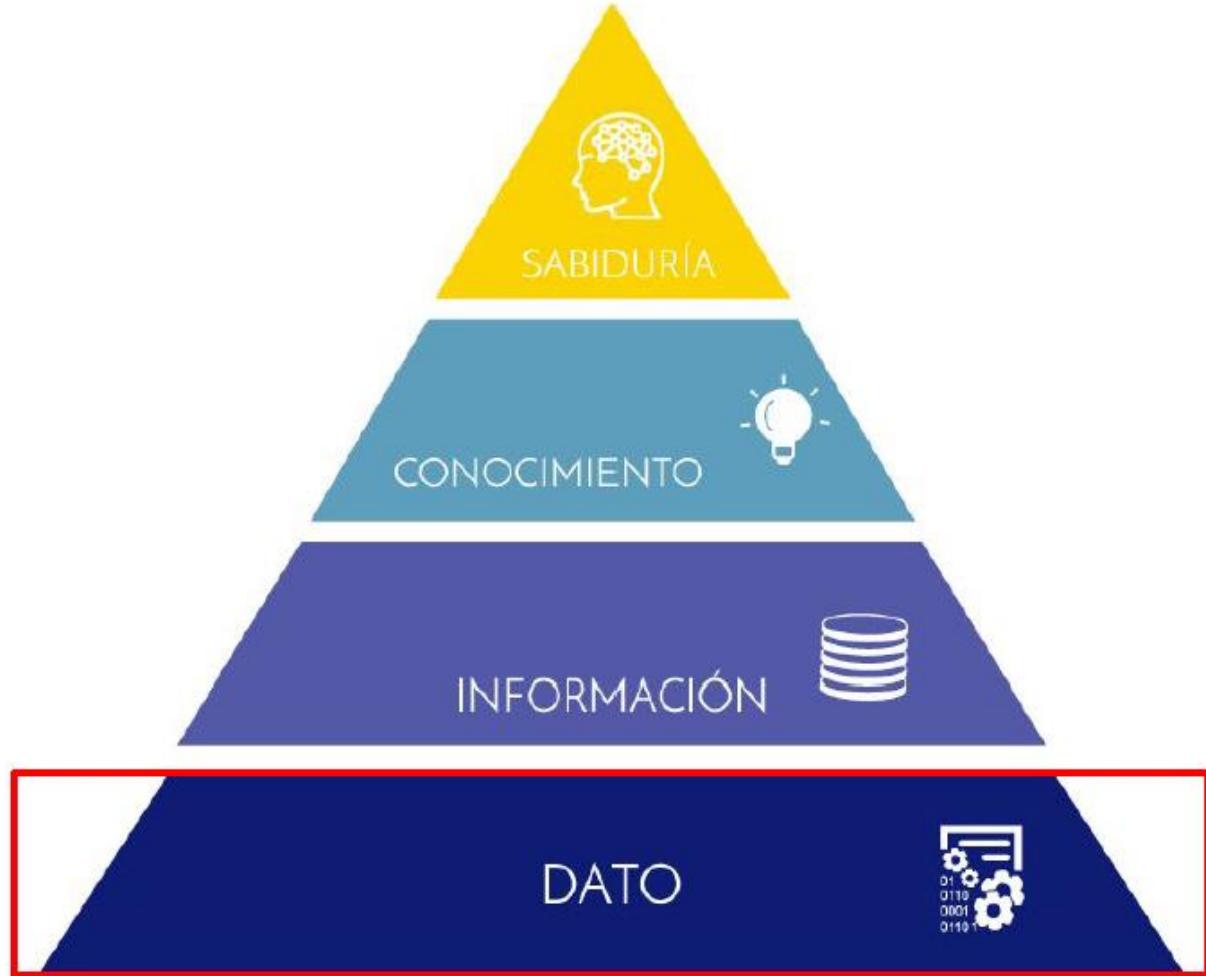
DATOS, INFORMACIÓN, CONOCIMIENTO

- Los **datos** son la mínima unidad de significado, elementos primarios de información que por sí solos son irrelevantes, que no dicen nada sobre el por qué de las cosas y no son orientativos para la acción.
- La **información** se puede definir como un conjunto de datos procesados e interrelacionados, que tienen un significado y por lo tanto son de utilidad para tomar decisiones.
- El **conocimiento** es el resultado de integrar los datos y la información con la experiencia, los valores y la personalidad, permitiendo su aplicación a la vida y a la toma de decisiones.



DATOS: PEQUEÑAS PINCELADAS

- Identificarlos
- Obtenerlos
- Transformarlos
- Almacenarlos
- Consumirlos
- Analizarlos



TIPOS DE DATOS

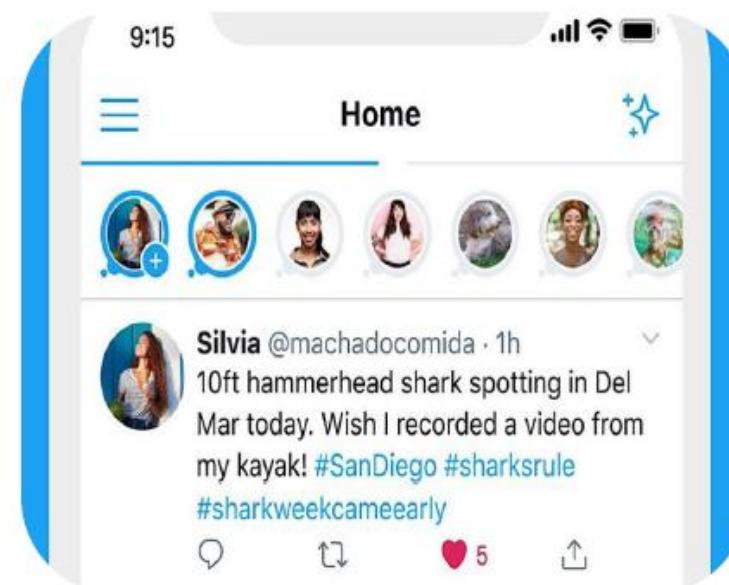
Estructurado

(Esquema: Tablas y campos). Rígidos.

Semiestructurado

Hay algo de estructura. CSV, JSON. Flexible.

No estructurado



KPI: SIGNIFICADO

El término KPI, siglas en inglés, de Key Performance Indicator, cuyo significado en castellano vendría a ser Indicador Clave de Desempeño o **Medidor de Desempeño**.

“El objetivo último de un KPI es ayudar a tomar mejores decisiones respecto al estado actual de **un proceso**, proyecto, estrategia o campaña y de esta forma, poder definir una línea de acción futura.”



QUÉ ES UN DASHBOARD



Se trata de una herramienta súper potente para obtener información de los datos y centralizar los **KPI** que necesitas para saber qué está pasando realmente **con tu negocio.**

EN RESUMEN...

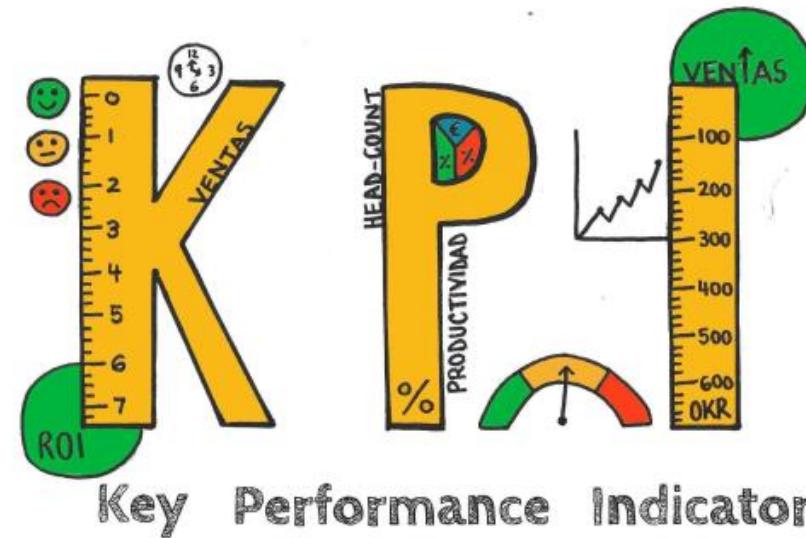
Dashboard = Cuadro de Mando que...

Ayuda a decidir porque de un vistazo sabemos “lo que pasa”

Esto implica que **ayuda ganando tiempo y esfuerzo** por lo que

El usuario **puede ser mejor** con una herramienta

Un Dashboard está compuesto de varias KPI = Métrica Clave. Cada una de ellas calcula un aspecto único dentro de la visión global.



Actividad:

- ¿Cuáles son los tipos de KPI?
- KPI's más utilizados
- KPI's y Ciencia de Datos



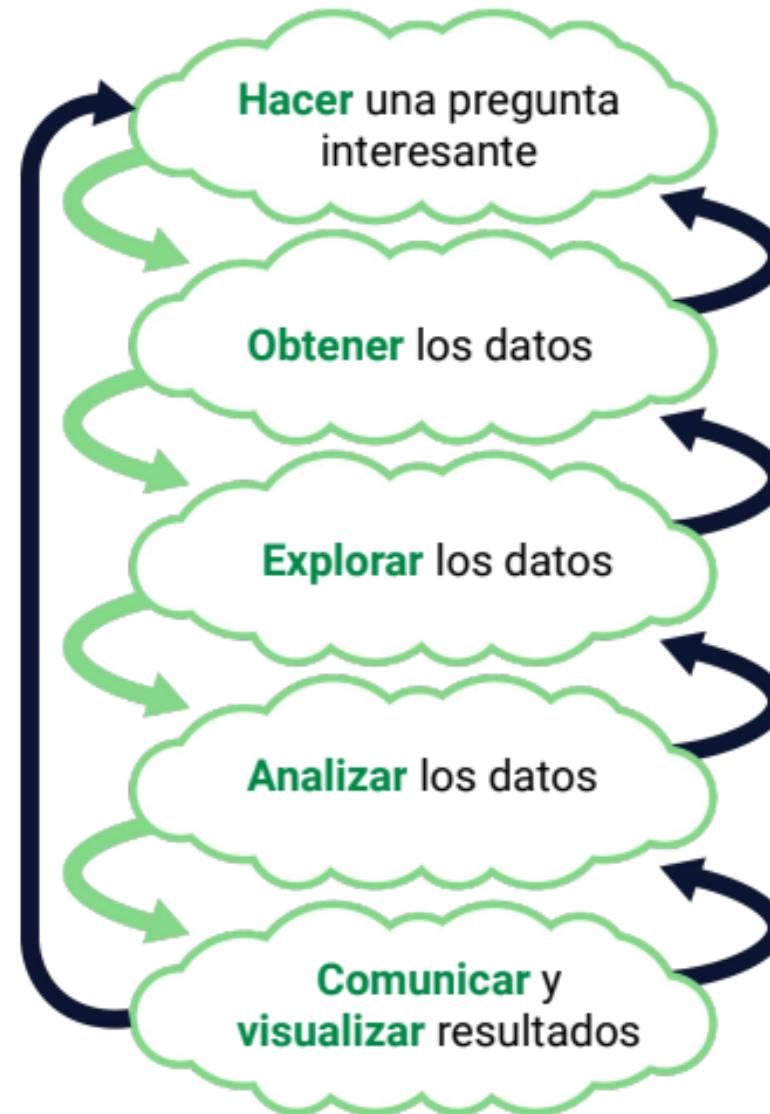
Actividad 1:

Preparar unas diapositivas en power point donde se resuma el contenido de los artículos sobre KPIs que tienes adjunto en tu Bloc de Notas. Aunque no vamos a exponer por ahora, asume que vas a realizar esta presentación a un público general.



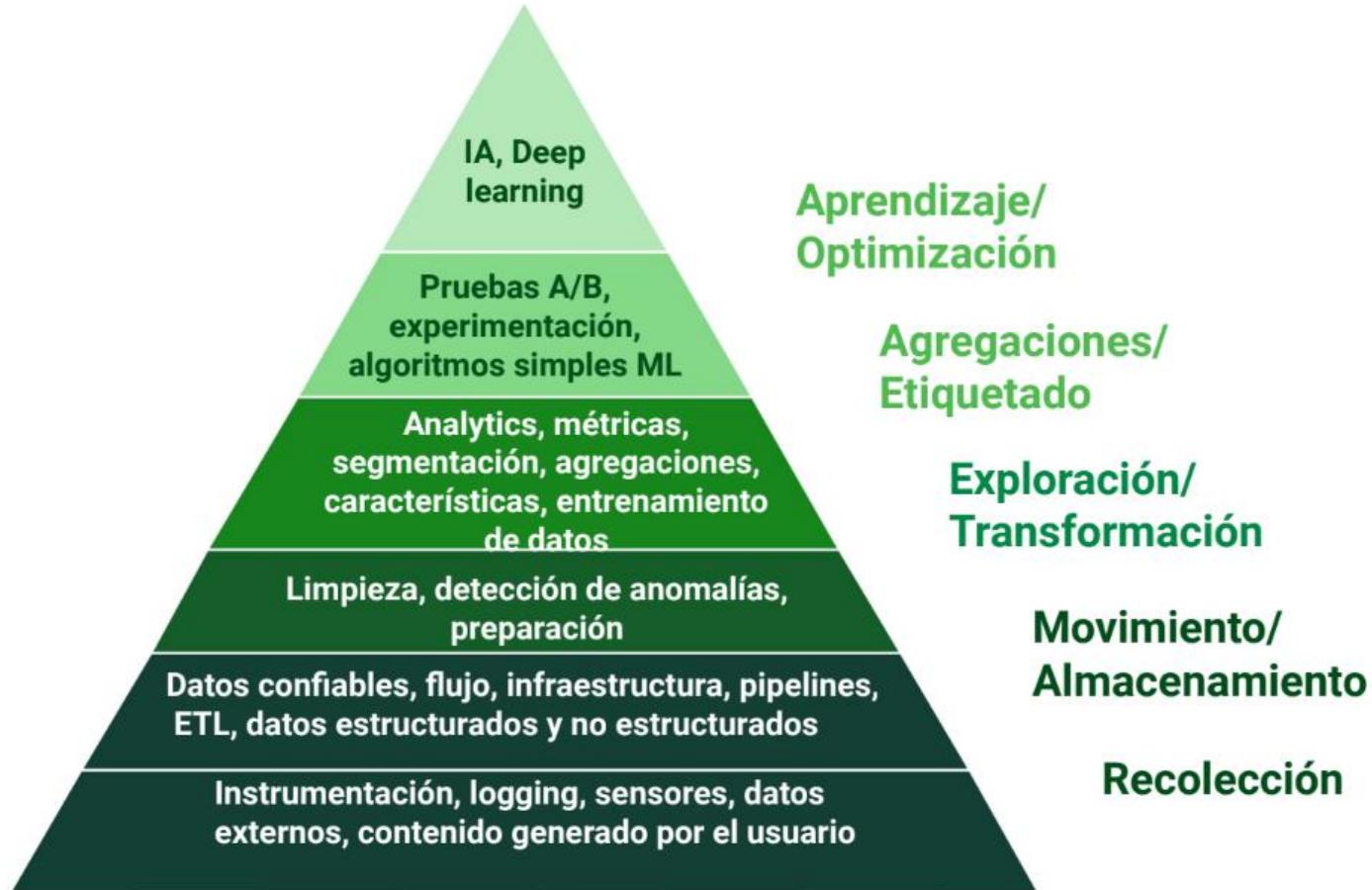
● Proceso de la ciencia de datos

- El proceso entre proyecto a proyecto cambia poco.
- Es el proceso del método científico llevado al uso de datos.





La jerarquía de necesidades de data science

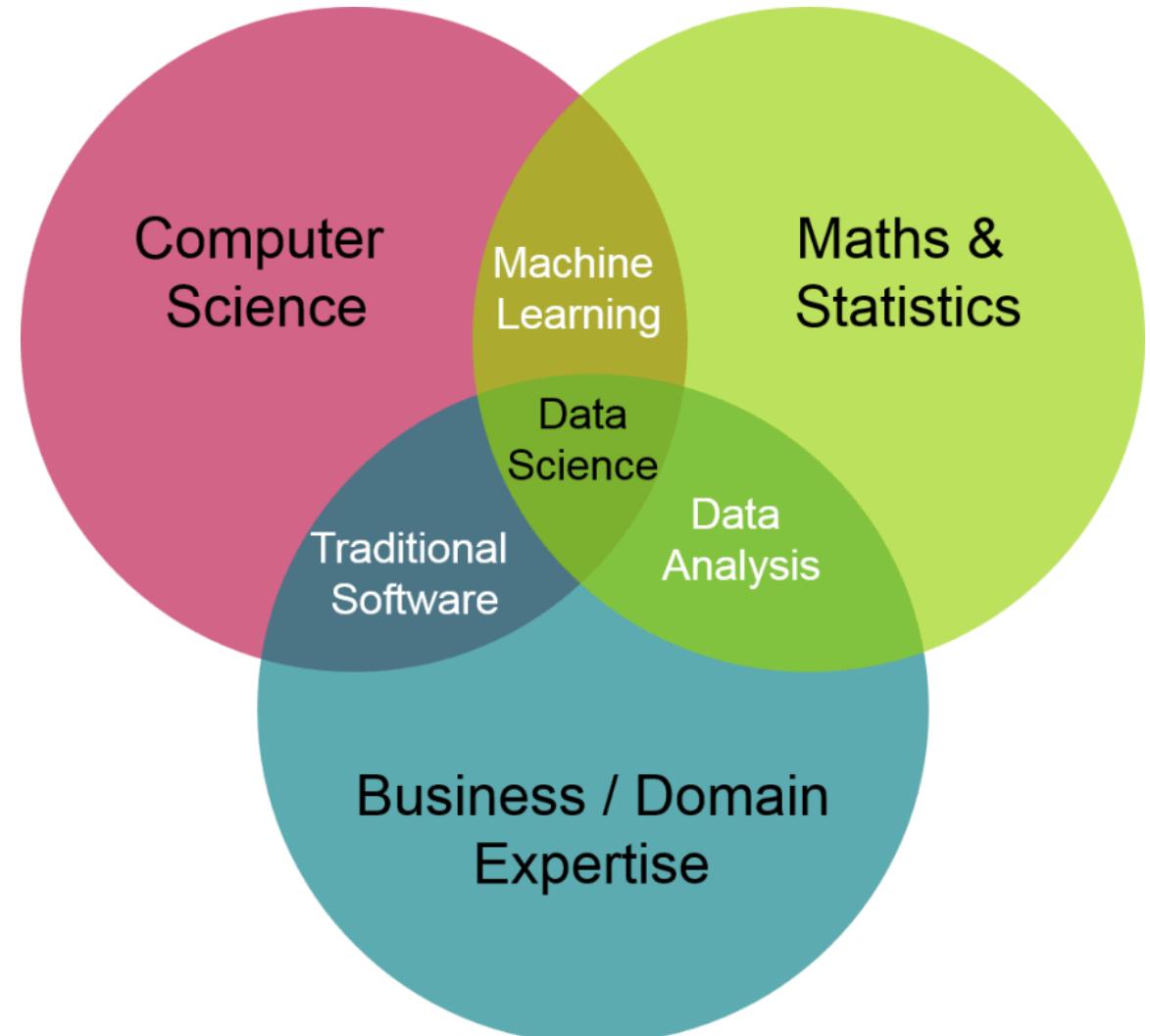


Referencia: 2. Data Science Hierarchy of needs (Monica Rogati – Hackernoon)

¿Qué es un Científico de Datos?

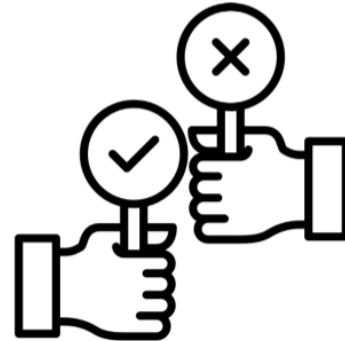
Es el profesional que, ante enormes bases de datos, aplica sobre ellas sus conocimientos en programación, matemáticas y estadística para recopilar, extraer y procesar información relevante que contienen.

Es alguien “que es mejor en estadística que cualquier programador, y mejor programador que cualquier estadístico” (Josh Wills).

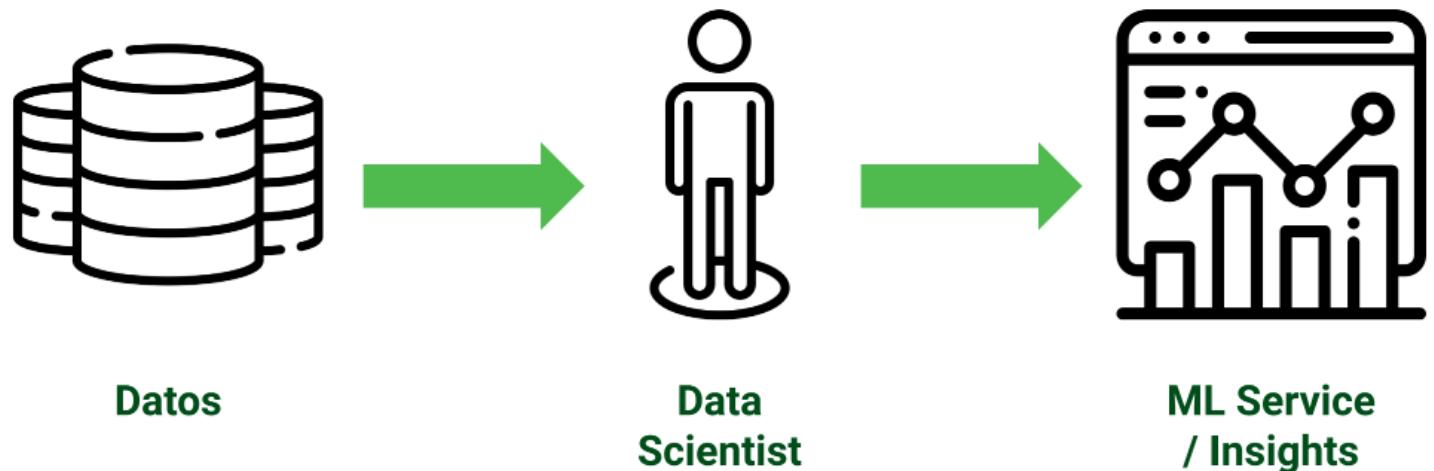


¿Qué hace un Científico de Datos?

Toma de decisiones
basadas en datos.



Incorporar datos a los
productos de software.



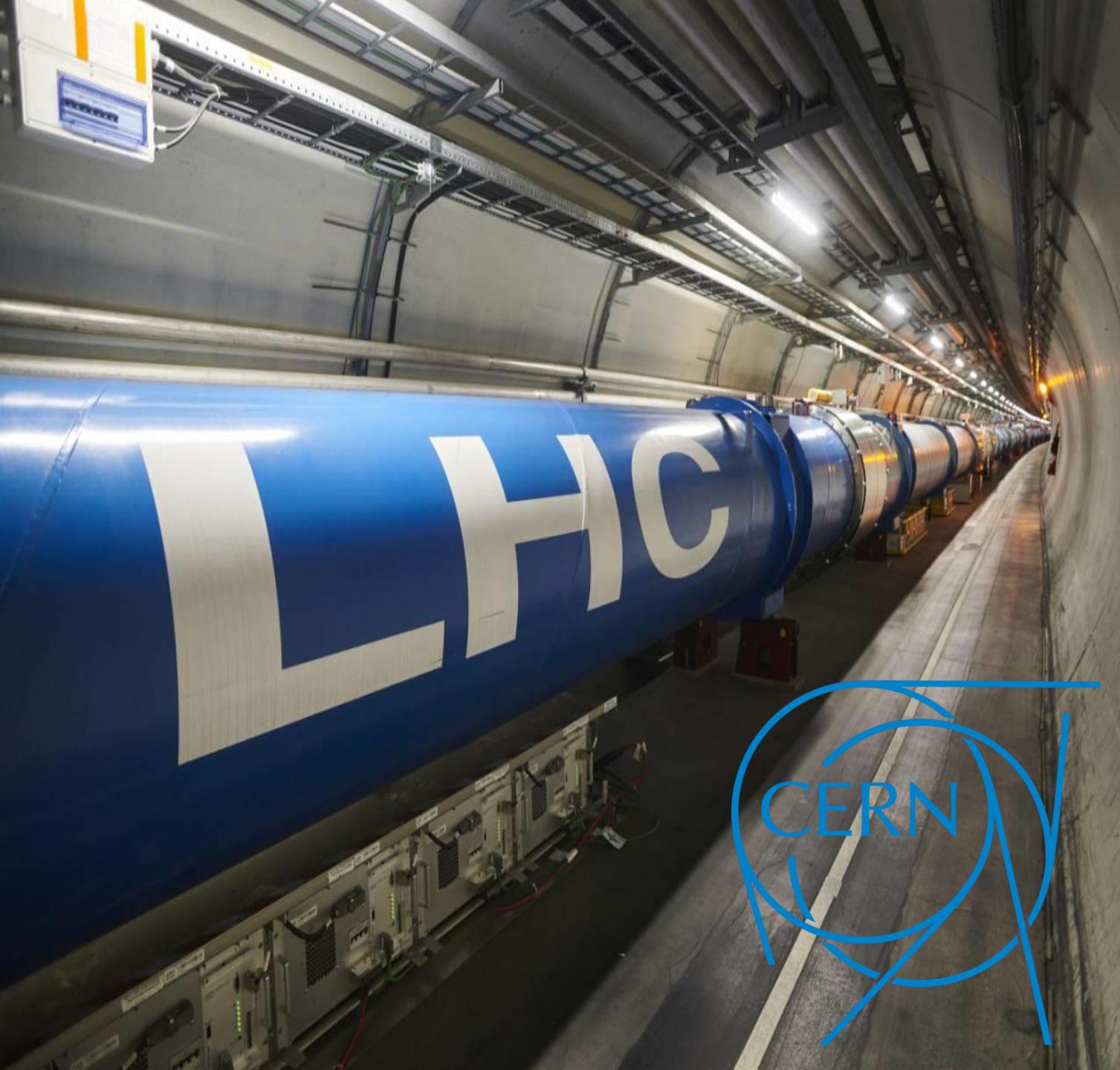
- | | | |
|---|--|--|
| <ul style="list-style-type: none">● Obtener, limpiar y procesar datos. | <ul style="list-style-type: none">● Diseñar y utilizar modelos de machine learning. | <ul style="list-style-type: none">● Crear reportes de información en tableros. |
| <ul style="list-style-type: none">● Monitorear la precisión de los datos. | | |
| <ul style="list-style-type: none">● Incorporar datos a los productos. | <ul style="list-style-type: none">● Automatizar procesos de recolección y transformación de datos. | <p>Día a día de
un Científico
de Datos</p> |

Ciencia de Datos en el contexto Big Data

¿Qué es Big Data?

Precursor del Big Data:
CERN





"El CERN ya desde hace mas de 20 años estaba aplicando Big Data en sus investigaciones."

Manuel Martín- Data Scientist-CERN

Cada año, el CERN produce 30 petabytes de nueva información, acumulando ya unos 250 Pb de datos en sus centros de datos, sobre los que se realizan unas dos millones de tareas cada día. Sólo en el CERN Logging Service (una herramienta que colecta y filtra los datos de todos los sensores) se generan 250 Gb de nuevos datos al día.

Y eso sólo de los detectores principales, ya que existen otros experimentos secundarios en el LHC. Al no operar todos los días del año, genera una media de 200 o 300 TB de datos; un volumen complicado -pero factible- de manejar hoy en día.

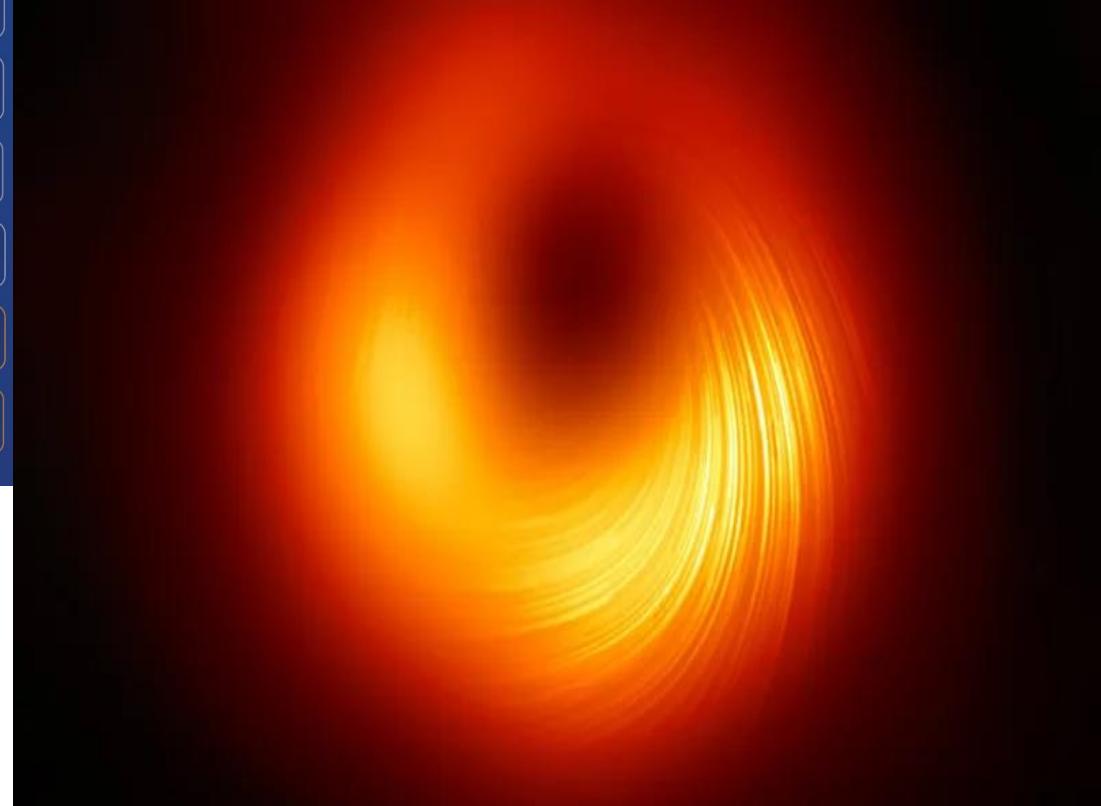
El problema es que el LHC entró en operación en 2008, cuando Big Data era un concepto muy novedoso, por lo que hubo mucho desarrollo de tecnología "sobre la marcha". No es la primera vez, ya que Internet mismo nació en el CERN, con la World Wide Web.

Event Horizon Telescope (EHT)

A Global Network of Radio Telescopes



Para poder generar esta imagen fue necesario un conjunto de 8 radio telescopios alrededor de mundo que se sincronizaron para obtener la información en tiempo real necesaria para generar el modelo que hemos podido ver.





En total, se transportó media tonelada de soportes de almacenamiento, que fueron procesados y analizados hasta generar la conocida imagen de menos de 1 MB.

A lo largo de las observaciones, cada telescopio generó unos 700 TB de datos, lo que resultó en un total de 5 PB de datos dispersos por tres continentes. El reto era combinar toda esta información en un solo lugar para su análisis.

Al contrario que en el LHC, no existía la infraestructura necesaria para transferencia de datos a ese nivel, ni merecía la pena desarrollarla al ser un caso de uso puntual. Por tanto, lo que se decidió fue transportar físicamente los discos duros por vía aérea, marítima y terrestre (1000 discos). Uno de los radiotelescopios estaba situado en la Antártida, y hubo que esperar al verano para que el deshielo parcial permitiera tener acceso físico a sus discos duros.

Ciencia de Datos en el contexto Big Data

¿Qué es Big Data?

Big Data hace referencia a un volumen masivo de datos que es tan grande que es difícil de procesar utilizando tecnología tradicional. De modo que requiere el uso de tecnologías especiales (cloud por ejemplo)

En la mayoría de los escenarios empresariales, el volumen de datos es demasiado grande o se mueve demasiado rápido o supera la capacidad de procesamiento actual.

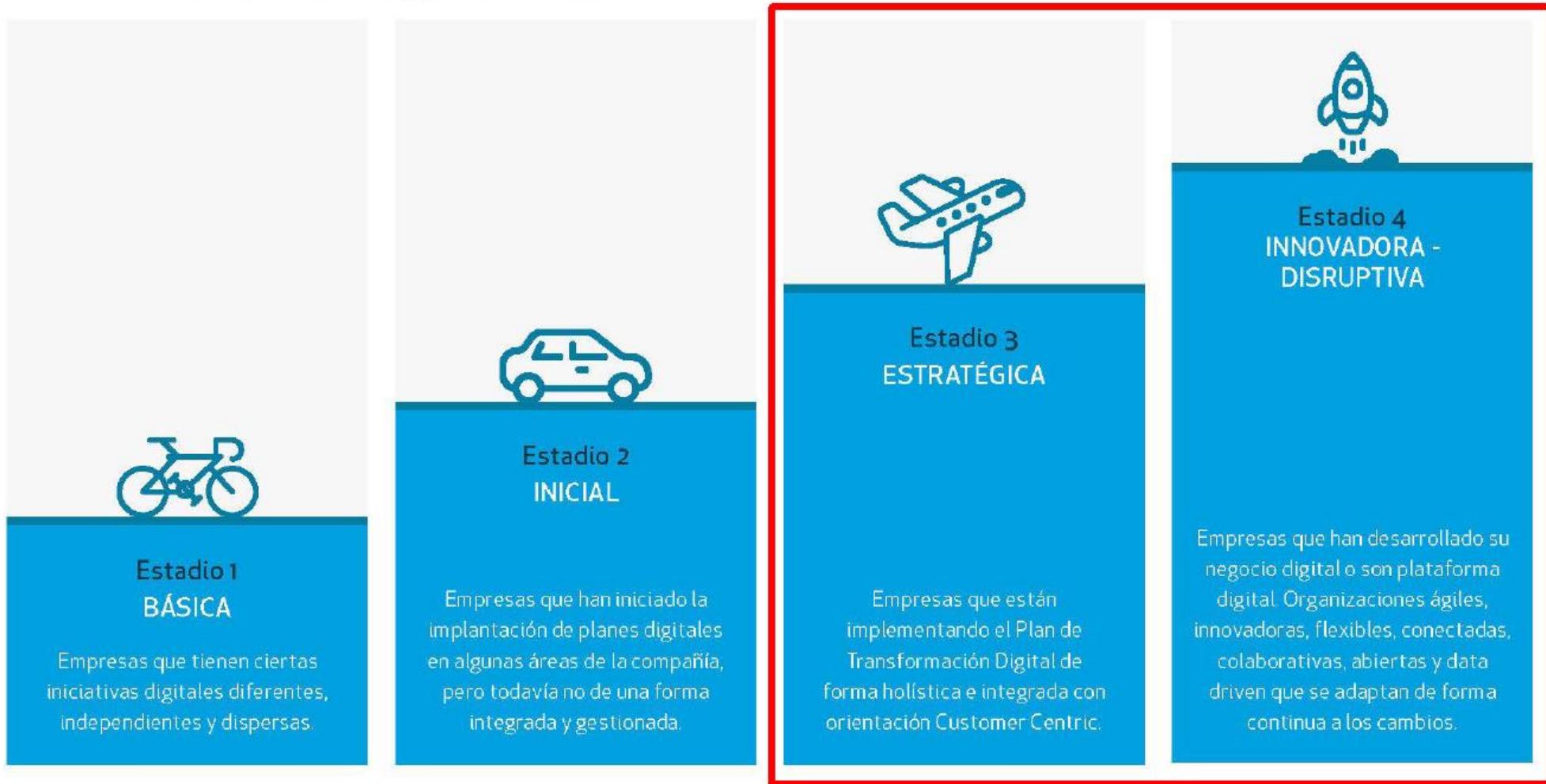


Madurez Digital Big Data

Entendemos la madurez digital como el estado en el que se encuentra una empresa, desde el punto de vista de hardware, software e incluso factor cultural. Es decir, qué métodos, herramientas y recursos ha desarrollado una empresa para implementar la transformación digital en su negocio.

Antes de hablar de Big Data debemos conocer el escenario de una empresa... No siempre tiene sentido hablar de estas soluciones.

- ¿Dónde se ubica una organización?



NIVELES DE MADUREZ DEL BIG DATA

Ser autosuficiente a nivel de recopilación y análisis de datos.

Aumentar la colaboración y la compartición de datos en todos los niveles empresariales.

Nivel 5

Análisis y datos como servicio



Nivel 4

Aplicación en empresa

Integración de metadatos, del departamento de calidad y gestión alrededor de la gestión de datos.

Aplicación de soluciones predictivas aplicadas a operaciones de negocio.

Análisis estructurados de datos.

Análisis predictivos aplicados al Big Data.

Nivel 3

Aplicación en negocio



Nivel 2

Aplicación técnica

Utilización del Big Data principalmente para el almacenamiento.

Primeros análisis de Datos.

Contemplar el uso de datos.

Primeros tests de implantación de recogida de datos.

Nivel 1

Infancia

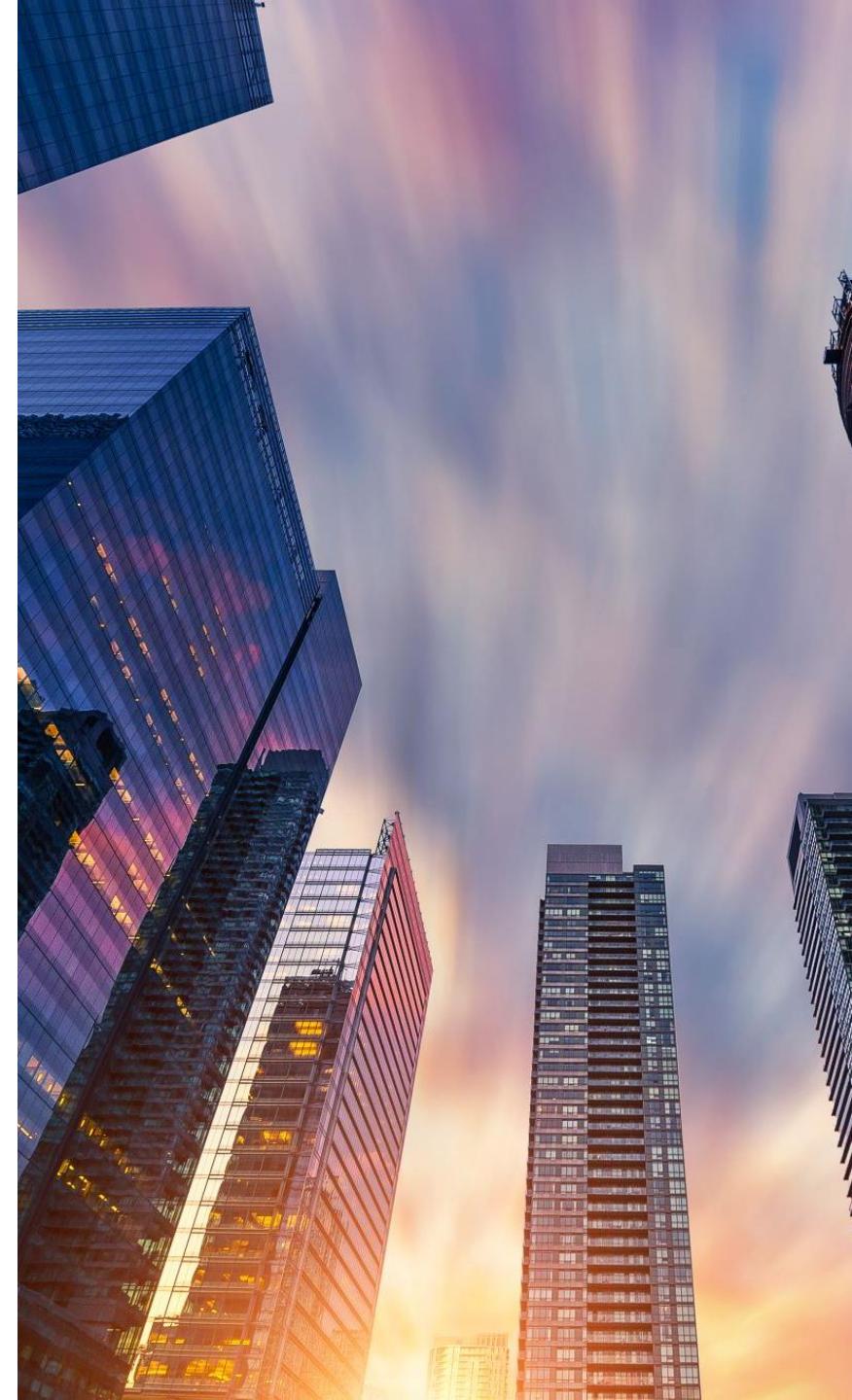


Actividad 2:

Exposiciones.

¿Temática por sorteo o libre elección?

- 1.- Digital Maturity: Definition, Models & Measuring (**Digital Maturity ≠ Digital Transformation**)
- 2.- Data Driven Maturity: Definition, Models & Measuring
- 3.- Big Data Maturity: Definition, Models & Measuring
- 4.- Data Science Maturity: Definition, Models & Measuring
- 5.- GenAI Maturity: Definition, Models & Measuring



La descripción de big data debe incluir las 3 V o las 5 V.

Inicialmente eran tres:

1. Volumen de datos: la gran cantidad de datos
2. Variedad de datos: tipos dispares, diferentes estructuras y formatos de datos
3. Velocidad de datos: qué tan rápido se agregan datos a los sistemas, se actualizan.

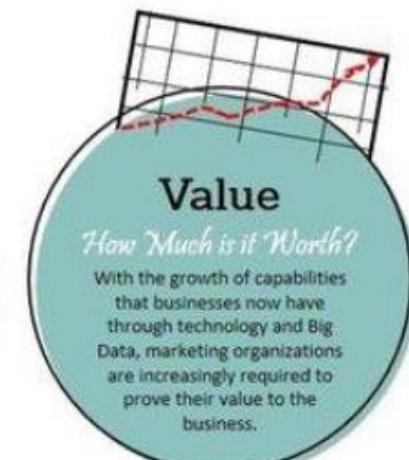
Luego se agregaron dos cualidades más para convertirlo en las 5 V de Big Data.

4. Valor: ¿Cuál es el retorno de la inversión para obtener estos datos?
5. Veracidad: ¿Cuál es la calidad y confiabilidad y de los datos?



LAS 4V

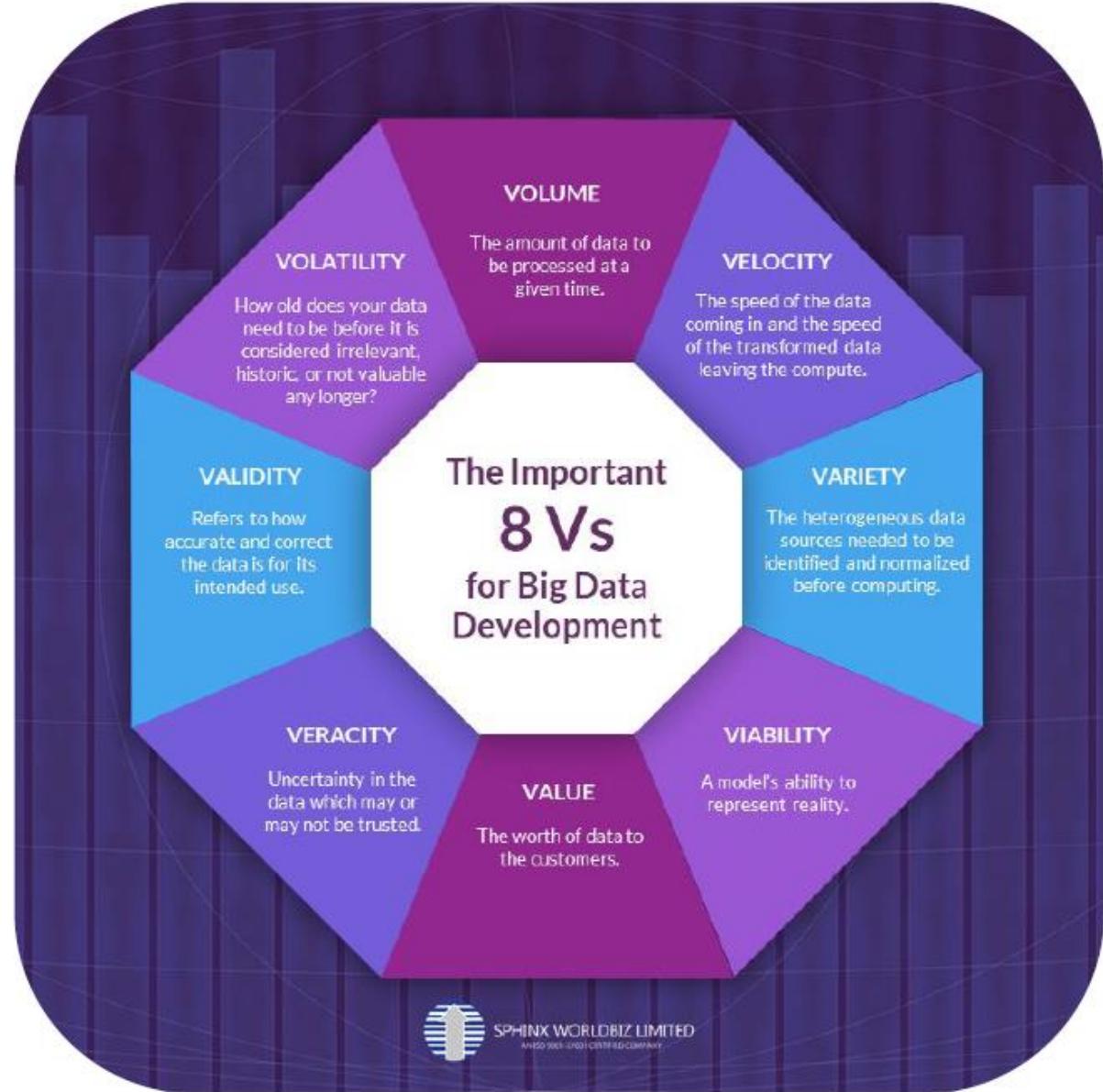
- VOLUMEN de información
- VELOCIDAD a la que se consume una solución
- VARIEDAD de información (diferentes formatos y tipos de datos)
- **VALOR:** Que valga para aprender o descubrir, analizar o decidir sobre un proceso.

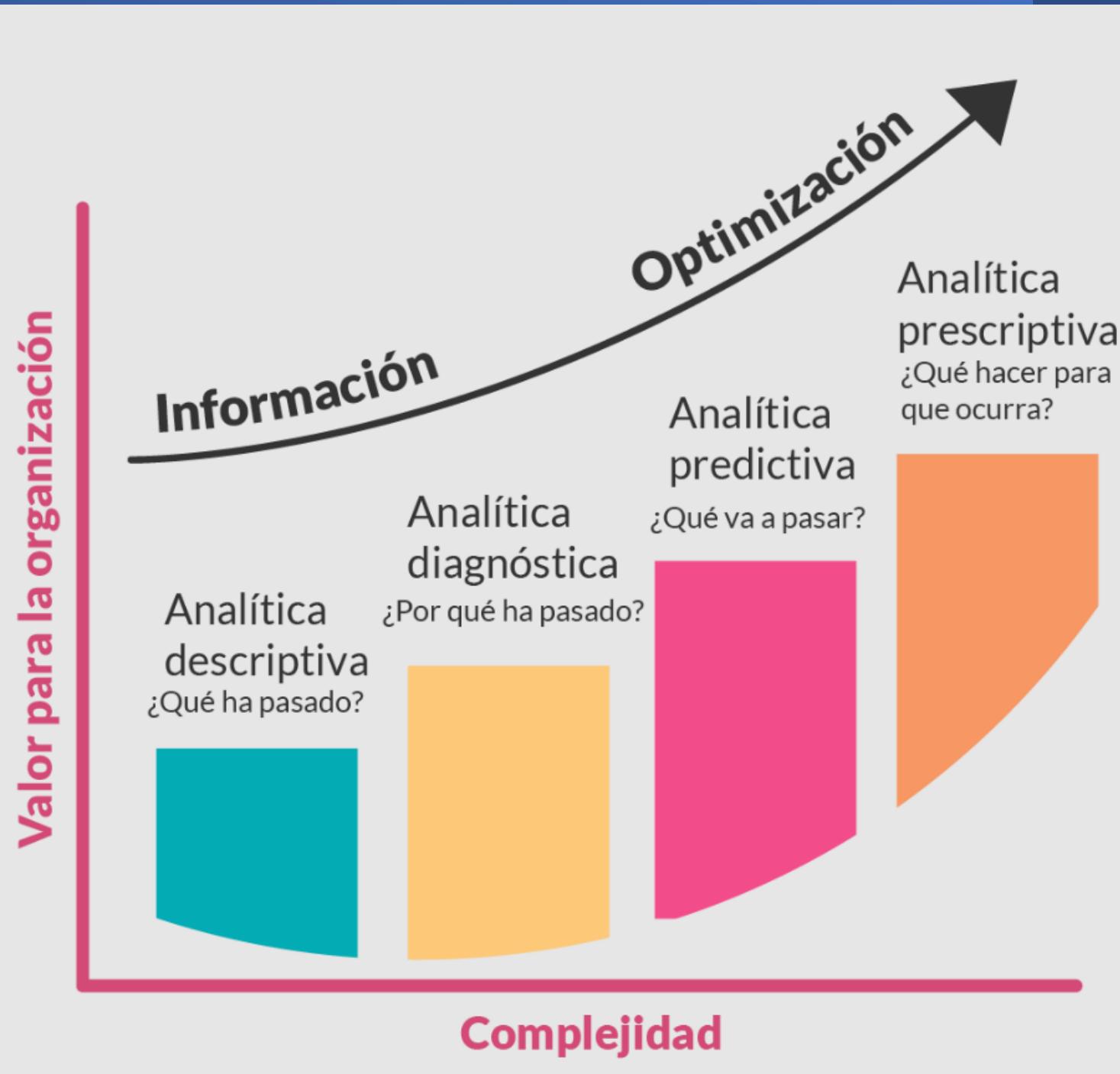


LAS 4V... u 8?

Algunos dicen que hay 8 V...

- **Viabilidad**
- **Veracidad**
- **Validez**
- **Volatilidad**





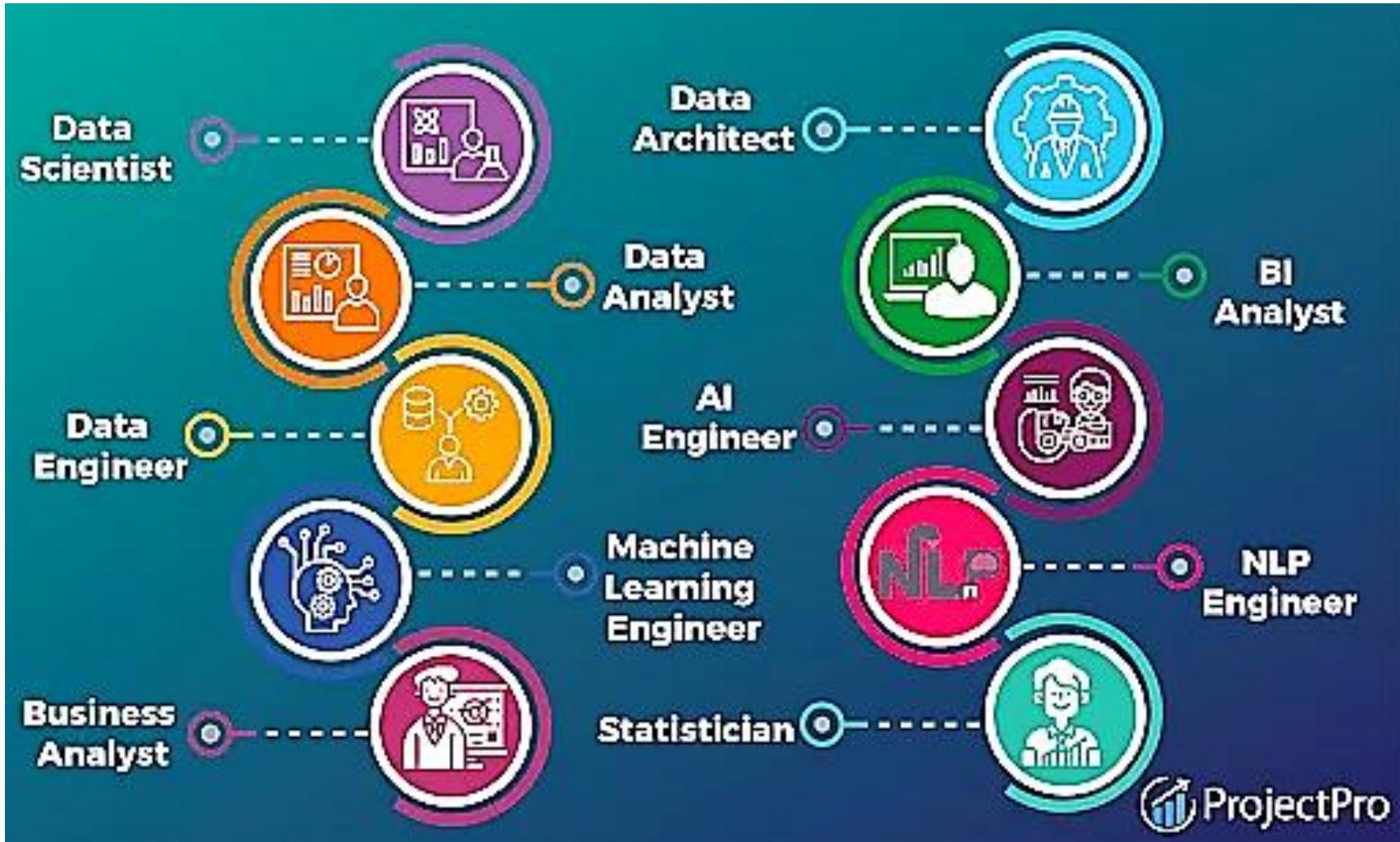


Procesamiento de Big Data

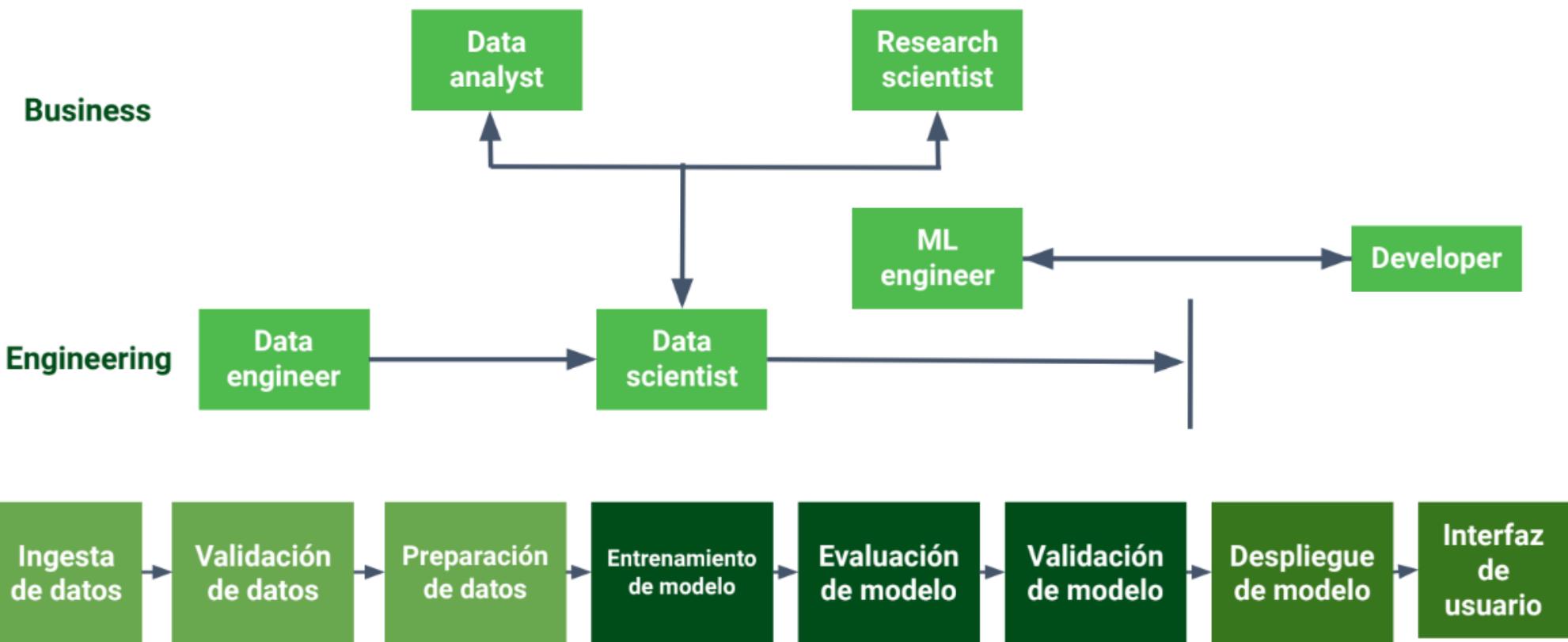
- Se procesa al dividirla en partes pequeñas en varias máquinas.
- Tecnologías como Spark, Hadoop y servicios de cómputo en la nube.



Ciencia de Datos en el contexto Big Data

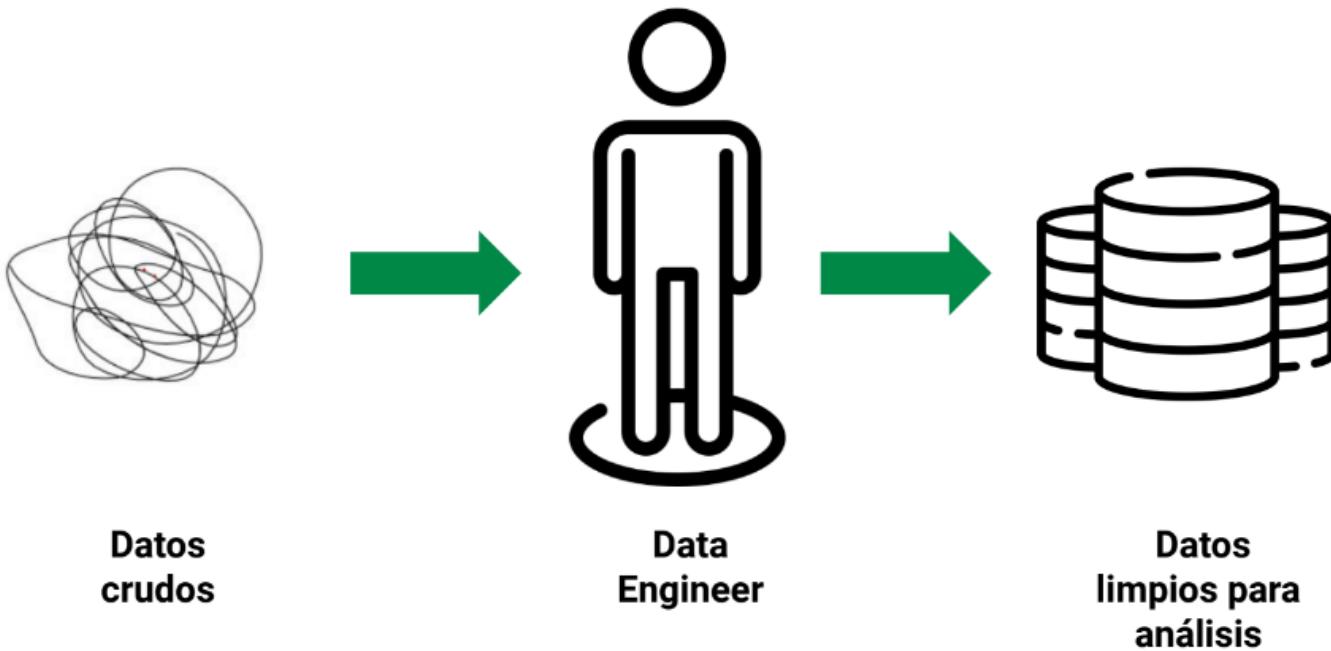


*Machine
Learning Design
Patterns,
Lakshmanan,
Robinson &
Munn (2021)*





¿Qué hace una Data Engineer?





¿Qué hace un Data Engineer?

Trabaja para que el equipo tenga datos para análisis.



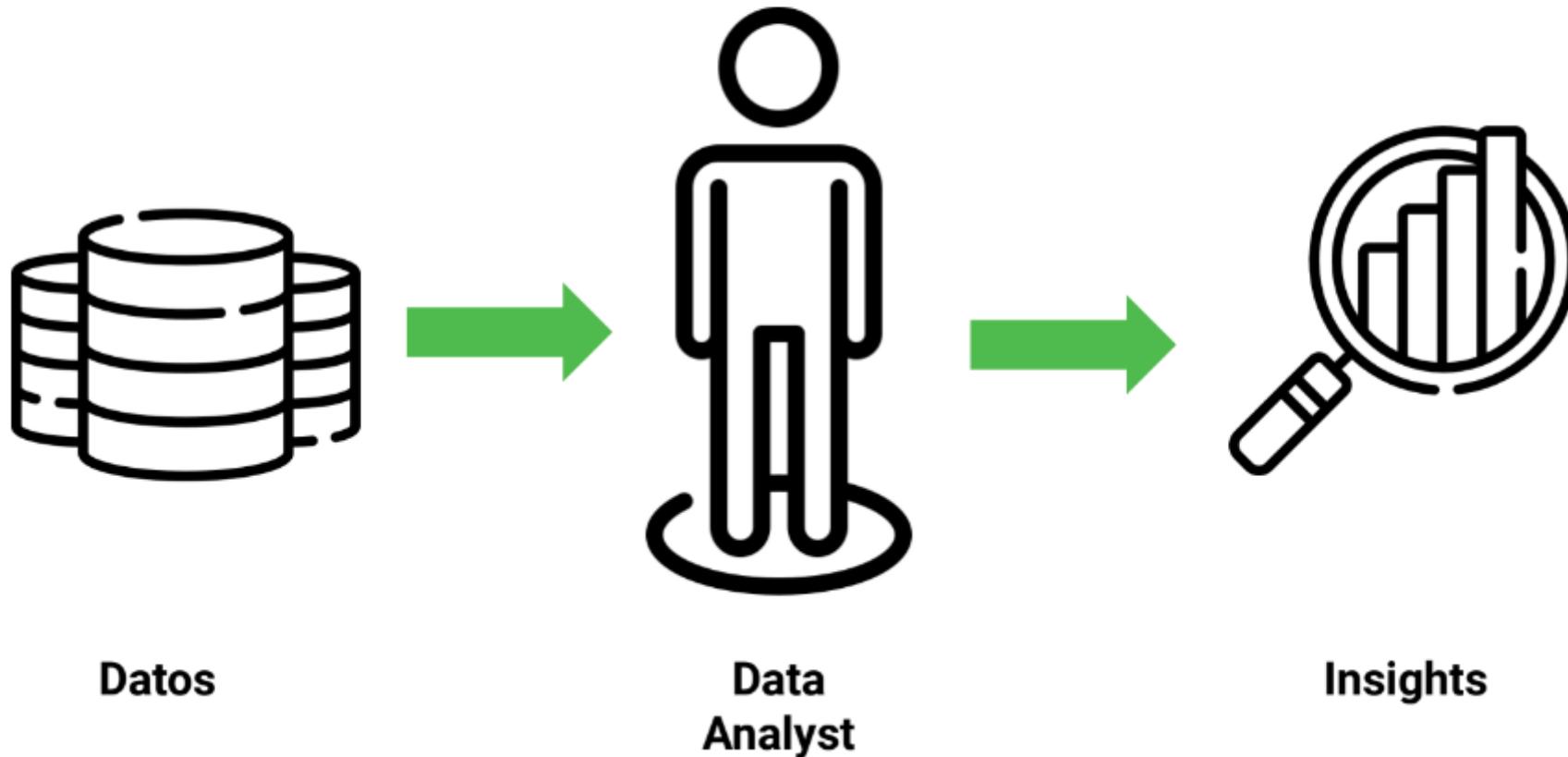
Crea pipelines ETL.



- Crear automatizaciones para ETL.
 - Data pipelines de ETL y bases de datos.
 - Transformar los datos para análisis.
-
- Extraer datos de diferentes fuentes.
 - Roles relacionados
 - Data Architect
 - Big Data Architect
 - Bases de datos especializadas para análisis.



¿Qué hace una Data Analyst?





• ¿Qué hace un Data Analyst?

Extraer datos recolectados.



Analizarlos y reportar resultados.



- **Limpiar y organizar los datos** para su análisis.
 - **Analizar los datos** para identificar patrones y tendencias.
-
- Comunicar los hallazgos en tableros o dashboards.



- **Identificar necesidades de información.**
- **Extraer datos de fuentes** con SQL o Python.

● Flujo de trabajo de Data Analyst



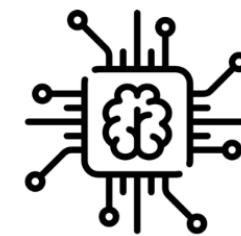
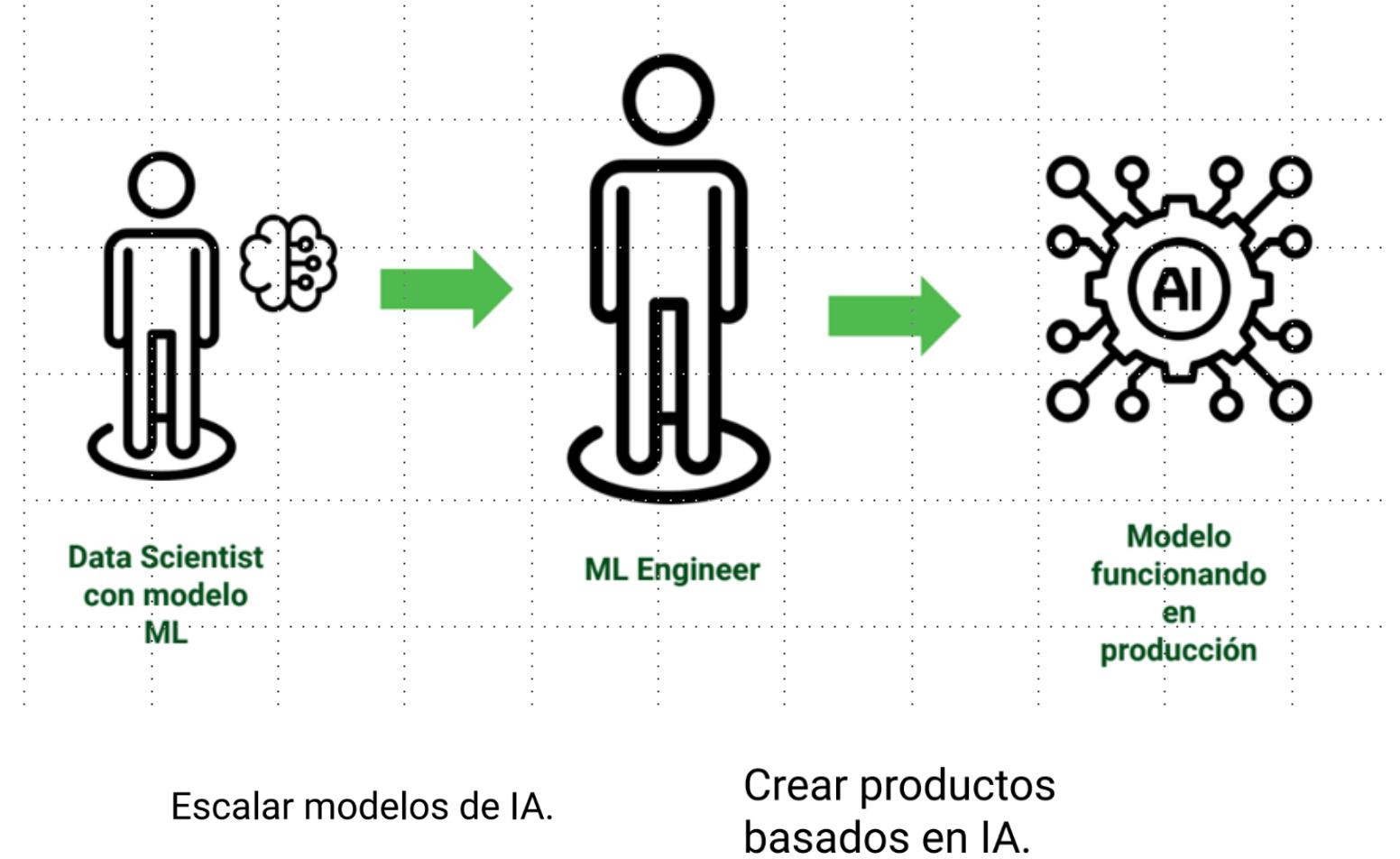


Roles relacionados

- Business Analyst
- Data visualization specialist



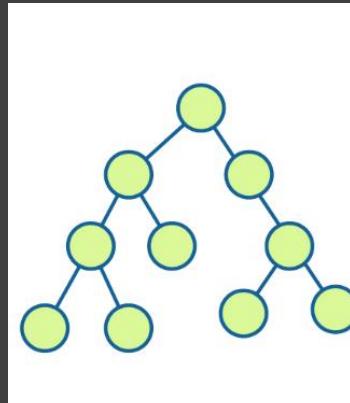
¿Qué hace un ML Engineer?



- **Monitorear el desempeño y funcionalidad** de los sistemas de machine learning.



Día a día de ML Engineer



- **Colaborar** con Data Scientists y otras áreas de ingeniería de software.

- Construir, escalar y robustecer sistemas de machine learning que **funcionen en producción**.

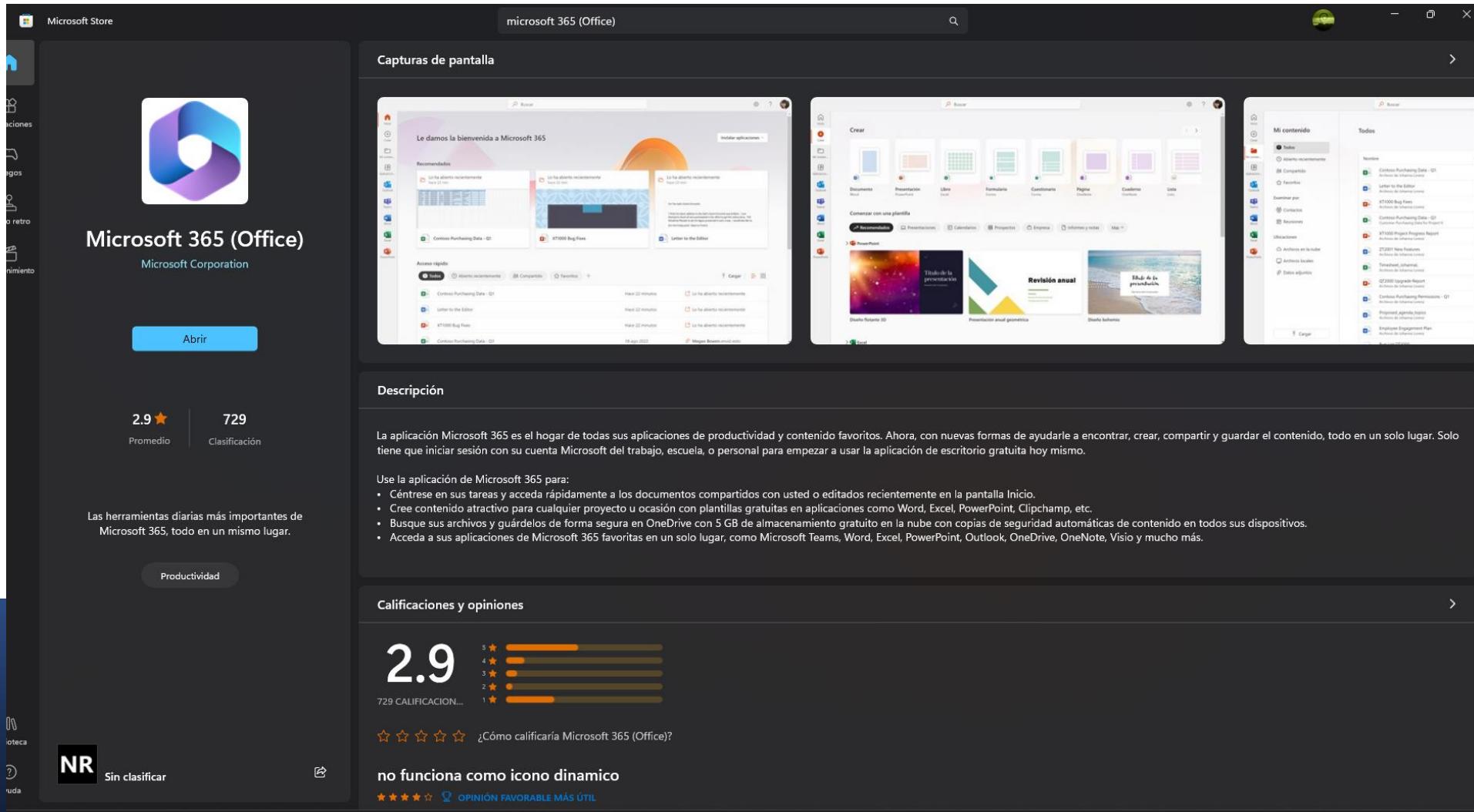
- Generar una **evaluación extensiva de métricas** de modelos de machine learning.

Proceso de machine learning



Actividad 2.

Instalar Microsoft 365 (Office) en el ordenador con las credenciales de Tajamar desde Microsoft Store



¿Que hace realmente un Científico de Datos?

- *De vuelta al científico de datos...*

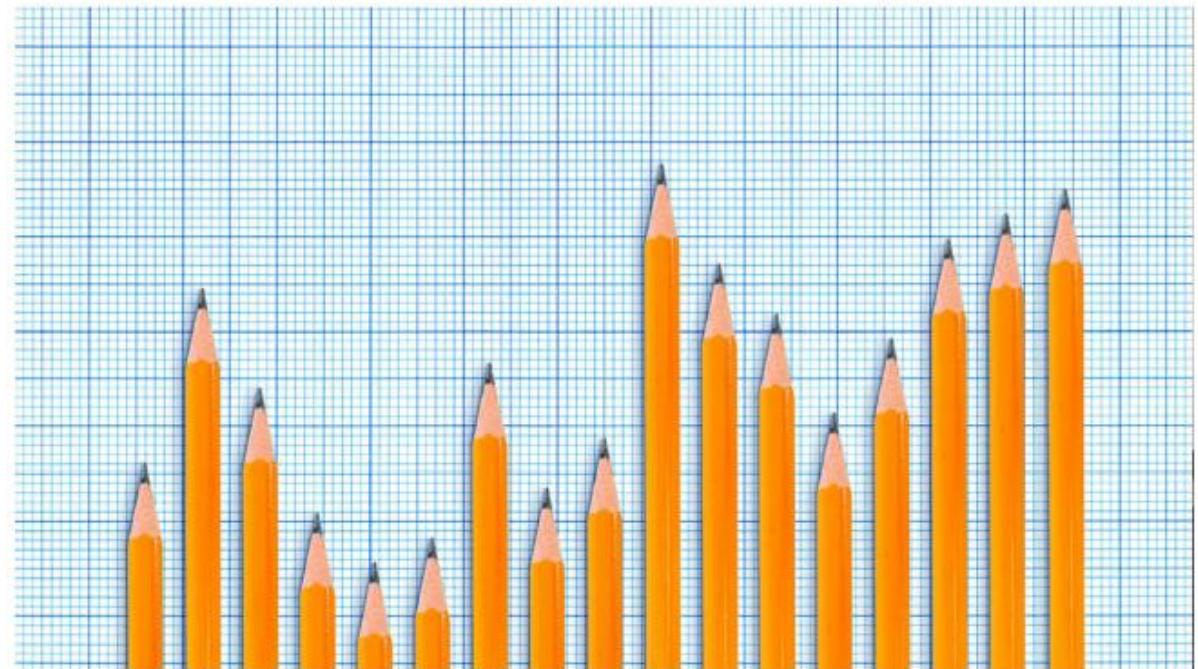
Harvard
Business
Review

Analytics And Data Science

What Data Scientists Really Do, According to 35 Data Scientists

by Hugo Bowne-Anderson

August 15, 2018



burakpekakcan/Getty Images



DATA SCIENTIST

AS RARE AS UNICORNS

- En primer lugar, los científicos de datos establecen una base de datos sólida para realizar análisis sólidos.
- Luego utilizan experimentos en línea, entre otros métodos, para lograr un crecimiento sostenible.
- Finalmente, crean canalizaciones de aprendizaje automático y productos de datos personalizados para comprender mejor su negocio y clientes para tomar mejores decisiones.

En otras palabras, en tecnología, la ciencia de datos se trata de infraestructura, pruebas, aprendizaje automático para la toma de decisiones y productos de datos.

Robert Chang



DATA SCIENTIST

AS RARE AS UNICORNS

En una conversación con Jonathan Nolis, un líder en ciencia de datos en el área de Seattle que ayuda a las empresas de Fortune 500, planteamos la pregunta:

"¿Qué habilidad es más importante para un científico de datos: la capacidad de usar los modelos de aprendizaje profundo más sofisticados o la capacidad para hacer buenas diapositivas de PowerPoint?

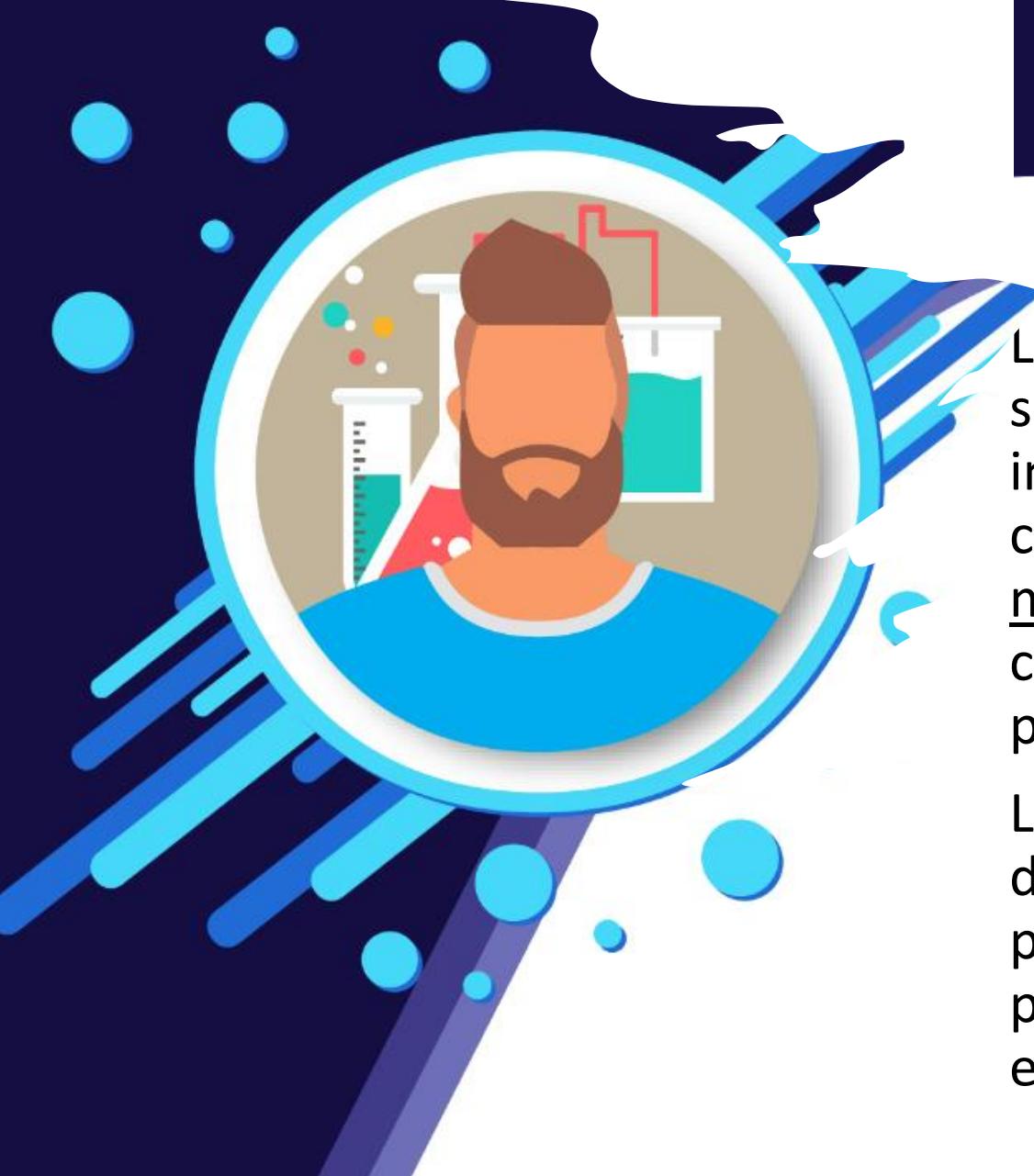
Defendió lo último, ya que comunicar los resultados sigue siendo una parte crítica del trabajo de datos.



DATA SCIENTIST

AS RARE AS UNICORNS

El 80 % del tiempo de un científico de datos se dedica simplemente a buscar, limpiar y organizar datos, dejando solo un 20 % para realizar el análisis.

A stylized illustration of a data scientist. The character has brown hair tied back, a beard, and is wearing a blue t-shirt. They are positioned inside a circular frame that looks like a magnifying glass or a target. In the background of the circle, there are laboratory glassware like a test tube and a beaker containing green liquid, along with some colorful dots. The entire illustration is set against a dark blue background with white bubbles.

DATA SCIENTIST

AS RARE AS UNICORNS

Las habilidades clave para los científicos de datos no son las habilidades para construir y usar infraestructuras de aprendizaje automático. En cambio, son las habilidades para aprender sobre la marcha y comunicarse bien para responder preguntas comerciales, explicando resultados complejos a las partes interesadas no técnicas.

Los aspirantes a científicos de datos, entonces, deberían centrarse menos en las técnicas que en las preguntas. Las nuevas técnicas van y vienen, pero el pensamiento crítico y las habilidades cuantitativas específicas del dominio seguirán siendo demandadas.



DATA SCIENTIST

AS RARE AS UNICORNS

Jonathan Nolis divide la ciencia de datos en tres componentes:

- (1) Inteligencia comercial (Business Intelligence), que se trata esencialmente de "tomar los datos que tiene la empresa y mostrarlos a las personas adecuadas" en forma de tableros e informes
- (2) ciencia de decisiones (Decision Science), que se trata de "tomar datos y usarlos para ayudar a una empresa a tomar una decisión"; y
- (3) aprendizaje automático, que se trata de "cómo podemos tomar modelos de ciencia de datos y ponerlos en producción de forma continua".

Aunque muchos científicos de datos en activo actualmente son generalistas y hacen las tres cosas, estamos viendo cómo surgen distintas trayectorias profesionales, como en el caso de los ingenieros de aprendizaje automático.

¿Por qué se considera un perfil unicornio?

DATA SCIENTIST

MUST-HAVE SKILLS

MATH & STATISTICS

- Machine Learning
- Statistical Modeling
- Exploratory Analysis
- Clustering
- Regression Analysis

DOMAIN KNOWLEDGE & SOFT SKILLS

- Inclination towards business operations
- Keen on working with data
- Problem solver
- Strategic, proactive, and cooperative
- Interested in hacking

PROGRAMMING & DATABASE

- Computer Science Fundamentals
- Database Management System
- Data Visualization
- Python
- Big Data

COMMUNICATION & VISUALIZATION

- Storytelling skills
- Convert data-based insights into decisions
- Collaborative with Sr. Management
- Knowledge of tools like Tableau
- Visual art design





Herramientas y tecnologías

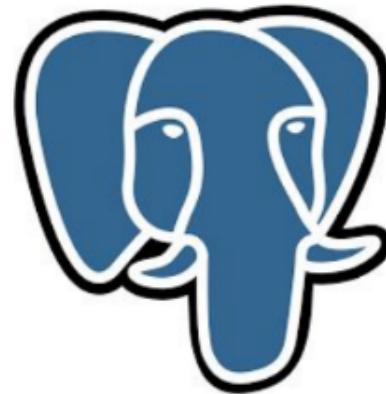
- Programación con Python o R (incluyendo POO).
- Jupyter Notebooks.
- Pandas, Numpy, Matplotlib.





Herramientas y tecnologías

- Algoritmos y librerías de machine learning como scikit-learn y TensorFlow.
- Bases de datos SQL y NoSQL.



cassandra



TensorFlow



Herramientas y tecnologías

- Software de visualización de datos como Power BI y Tableau.
- Excel y Google Sheets.



Registrarse en
GitHub
utilizando el
correo de
Tajamar.



GitHub



Visual Studio Code



Instalar y registrarse con la cuenta de GitHub que creasteis en el paso anterior.

Registrarse con la cuenta de GitHub o con la cuenta de Tajamar.

Crear una cuenta de Google y luego habilitar Google Colab



git



Herramientas de consulta.



towards
data science



[Registrarse](#)

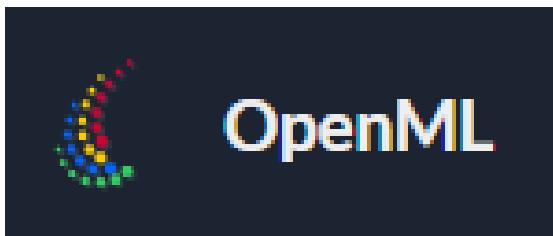


Papers With Code
Este es de nivel avanzado

[Registrarse](#)

Usar con mucho cuidado,
verificar fuentes.

Datos Abiertos para Prácticas



Open Data
on AWS Data
Exchange



datos.gob.es
reutiliza la información pública

Registrarse con la cuenta de GitHub o
con la cuenta de Tajamar.

datos abiertos

Google

Dataset Search

Search for Data Sets
Try [boston education data](#) or [weather site.noaa.gov](#)
Find out more about including your datasets in Dataset Search.

Registrarse e instalar en el ordenador con la cuenta de GitHub.

Anaconda Navigator

File Help

Connected to Cloud Connect ▾

Home Environments Learning Community

All applications on base (root) Channels

DataSpell Anaconda Notebooks CMD.exe Prompt JupyterLab Notebook Powershell Prompt

DS Anaconda Notebooks icon. Description: Cloud-hosted notebook service from Anaconda. Launch a preconfigured environment with hundreds of packages and store project files with persistent cloud storage. Launch button highlighted with a red box.

DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one user-friendly environment. Install button.

Anaconda Notebooks icon. Description: Cloud-hosted notebook service from Anaconda. Launch a preconfigured environment with hundreds of packages and store project files with persistent cloud storage. Launch button highlighted with a red box.

CMD.exe Prompt icon. Description: Run a cmd.exe terminal with your current environment from Navigator activated. Launch button highlighted with a red box.

JupyterLab icon. Description: An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture. Launch button highlighted with a red box.

Notebook icon. Description: Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis. Launch button highlighted with a red box.

Powershell Prompt icon. Description: Run a Powershell terminal with your current environment from Navigator activated. Launch button highlighted with a red box.

Qt Console Spyder VS Code Anaconda on AWS Graviton Datalore IBM watsonx

IP[y]: Spyder icon. Description: PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more. Launch button highlighted with a red box.

Qt Console icon. Description: PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more. Launch button highlighted with a red box.

Spyder icon. Description: Scientific Python Development Environment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features. Launch button highlighted with a red box.

VS Code icon. Description: Streamlined code editor with support for development operations like debugging, task running and version control. Launch button highlighted with a red box.

Anaconda on AWS Graviton icon. Description: Running your Anaconda workloads on AWS Graviton-based processors could provide up to 40% better price performance.

Datalore icon. Description: Kick-start your data science projects in seconds in a pre-configured environment. Enjoy coding assistance for Python, SQL, and R in Jupyter notebooks and benefit from no-code automations. Use Datalore online for free. Launch button highlighted with a red box.

IBM watsonx icon. Description: IBM watsonx is an enterprise-ready AI platform including a data store, model builder, and AI model management and monitoring. Launch button highlighted with a red box.

Anaconda Toolbox Oracle Cloud Infrastructure Oracle Data Science Service Glueviz Orange 3 PowerShell_shortcut_miniconda PyCharm Professional

Anaconda Toolbox icon. Description: Supercharged local notebooks. Click the Toolbox tile to install. Read the Docs button.

Documentation Anaconda Blog

Twitter YouTube GitHub

console_shortcut_miniconda icon. Description: Multidimensional data visualization across files. Explore relationships within and among related datasets. Launch button highlighted with a red box.

Oracle Data Science Service icon. Description: OCI Data Science offers a machine learning platform to build, train, manage, and deploy your machine learning models on the cloud with your favorite open source.

Glueviz icon. Description: Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a tree toolbar. Launch button highlighted with a red box.

Orange 3 icon. Description: Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a tree toolbar.

PowerShell_shortcut_miniconda icon. Description: A Full-fledged IDE by JetBrains for both Scientific and Web Python development. Supports HTML, JS, and SQL.

PyCharm Professional icon. Description: A Full-fledged IDE by JetBrains for both Scientific and Web Python development. Supports HTML, JS, and SQL.

Iniciar sesión en **Anaconda Cloud** con la misma cuenta que usaste para **Anaconda Navigator**



Welcome to Anaconda Cloud



Introduction to Anaconda

Watch an introductory course on Anaconda Distribution, conda, and creating your first Python program.

[Start Learning ▶](#)



Code Online

Now featuring new AI-powered code generation, insights, and debugging!

Prefer to code in your browser? Start coding immediately with Anaconda Notebooks! No installation or configuration necessary.

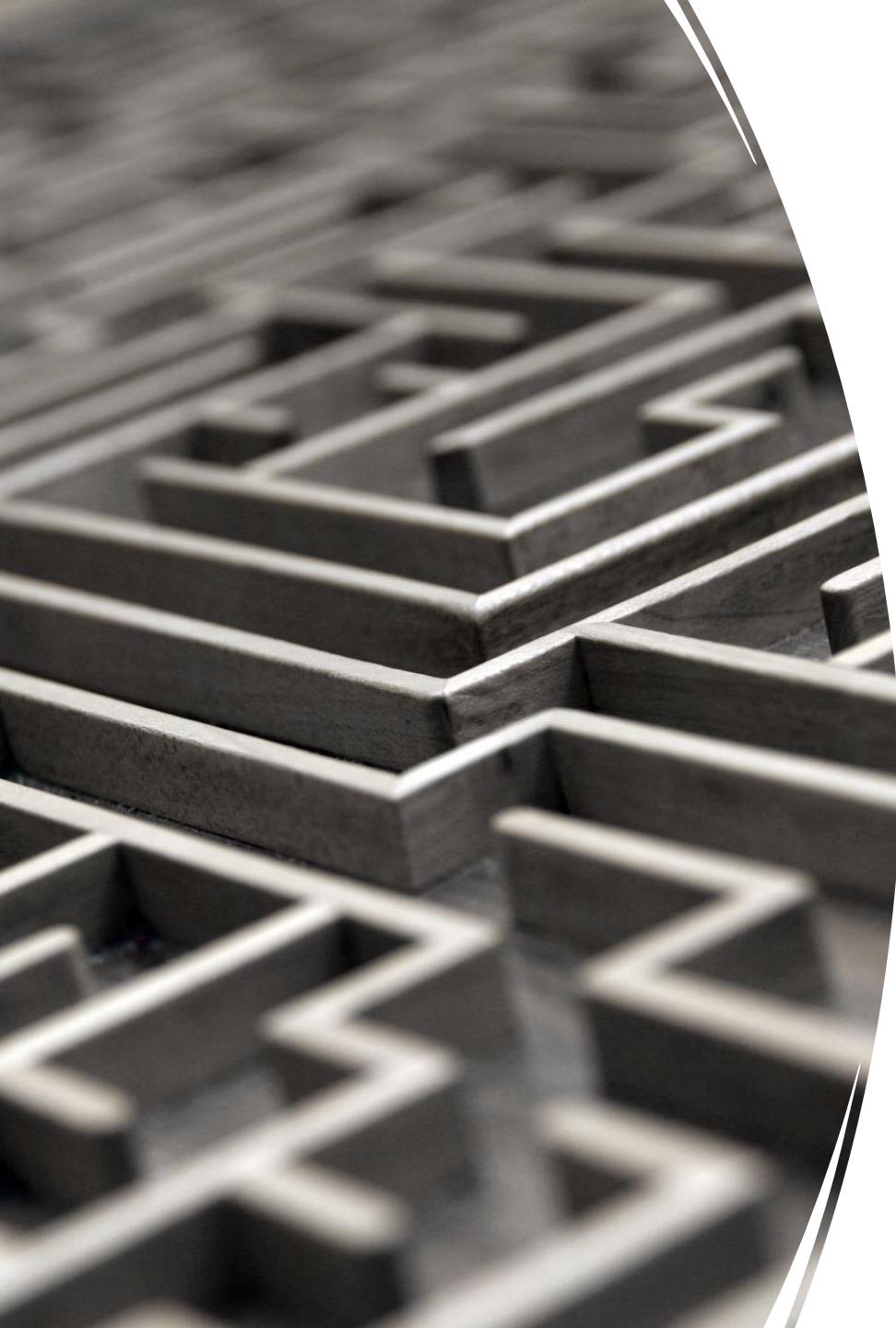
[See Sample Notebook ▶](#)



Anaconda Distribution

Get started with the most fundamental DS, AI, and ML packages. Easily manage applications, packages, and environments using Navigator instead of the command line.

[Install Distribution ▶](#)

A circular grayscale photograph showing a complex, three-dimensional metal labyrinth. The迷宫 (labyrinth) is constructed from dark, metallic bars forming a dense network of paths and dead ends. The perspective is from above, looking down into the center of the maze.

¿Qué perfiles están
buscando las
empresas?

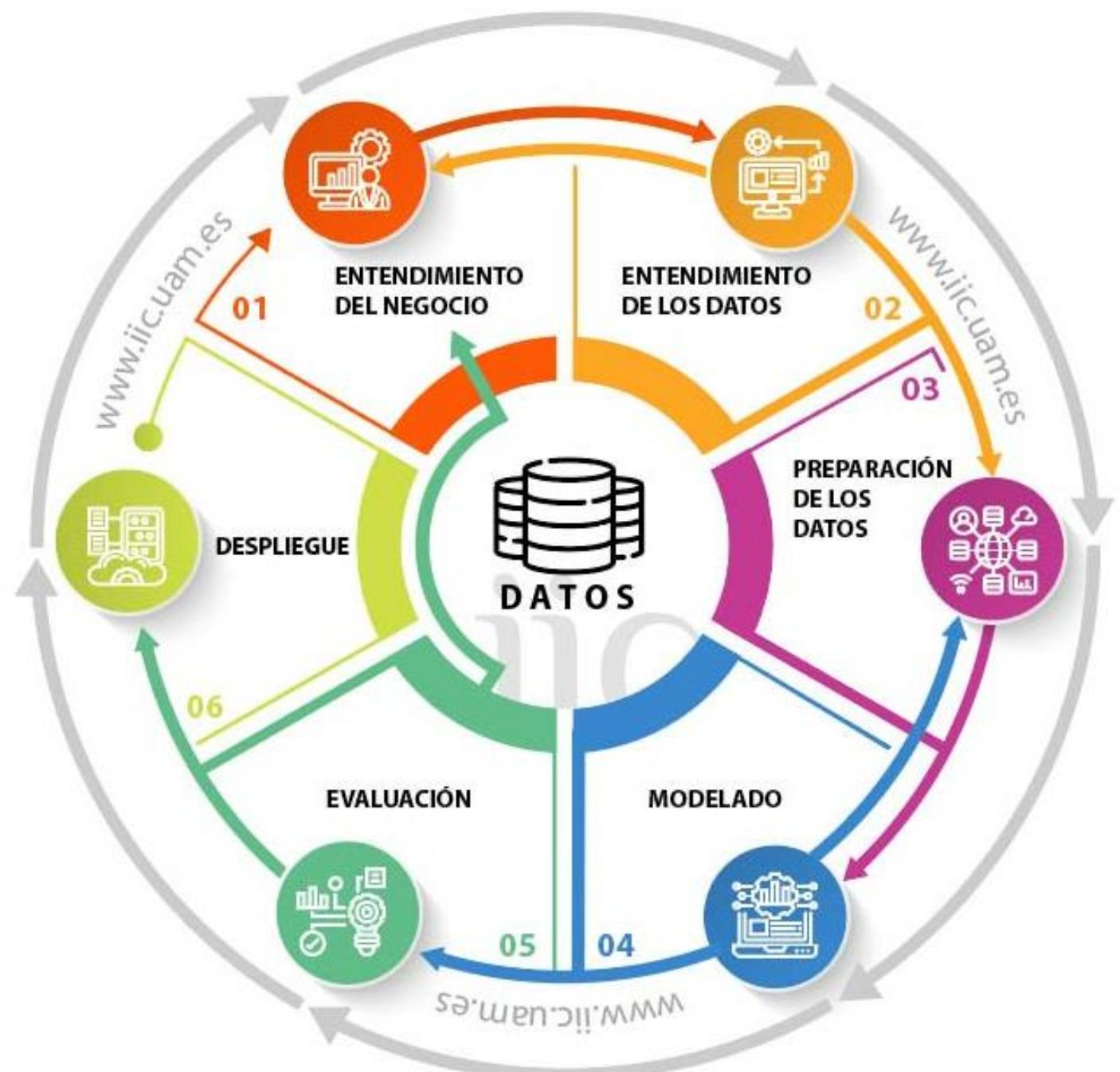
Actividad 3

Buscar mínimo 5 ofertas de empleo en LinkedIn de cada uno de los siguientes perfiles “Junior”: Data Analyst, Data Scientist, Data Engineer, Data Architect , Business Intelligence Analyst o sus similares, Inteligencia Artificial y sus similares.

Tomar apuntes de los requisitos/skills que se piden, en un documento de Word o Power Point y compartir en Teams en la pestaña de **Noticias, Eventos y Post**.



Modelo CRISP-DM



CRISP-DM: Cross Industry Standard Process for Data Mining

Business Understanding

Determine Business Objectives

Assess Situation

Determine Data Mining Goals.

Produce Project Plan

Data Understanding

Collect Initial Data

Describe Data

Explore Data

Verify Data Quality

Data Preparation

Select Data

Clean Data

Construct Data

Integrate Data

Format Data

Modeling

Select Modeling Techniques

Generate Test Design

Build Model

Assess Model

Evaluation

Evaluate Results

Review Process

Determine Next Steps

Deployment

Plan Deployment

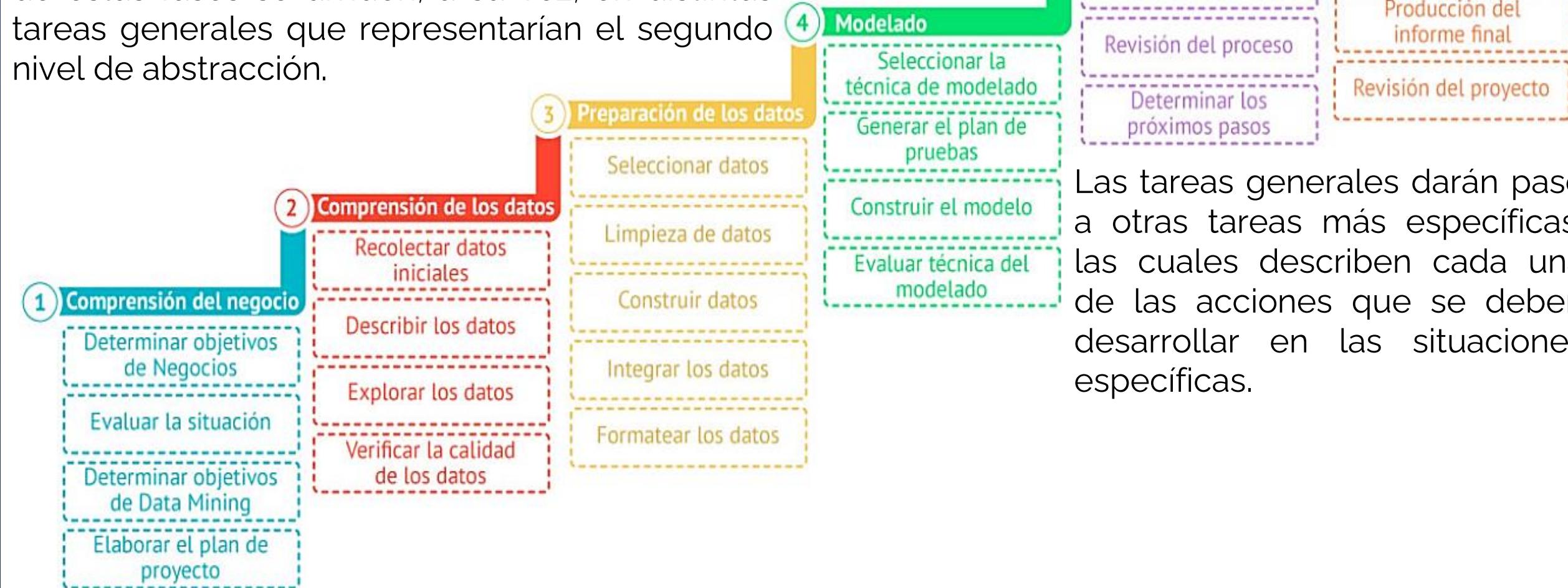
Plan Monitoring and Maintenance

Produce Final Report

Review Project

Fases y tareas de la metodología CRISP-DM

Considerando el nivel más general, **el proceso se organiza en seis fases principales**. Cada una de estas fases se dividen, a su vez, en distintas tareas generales que representarían el segundo nivel de abstracción.



Las tareas generales darán paso a otras tareas más específicas, las cuales describen cada una de las acciones que se deben desarrollar en las situaciones específicas.

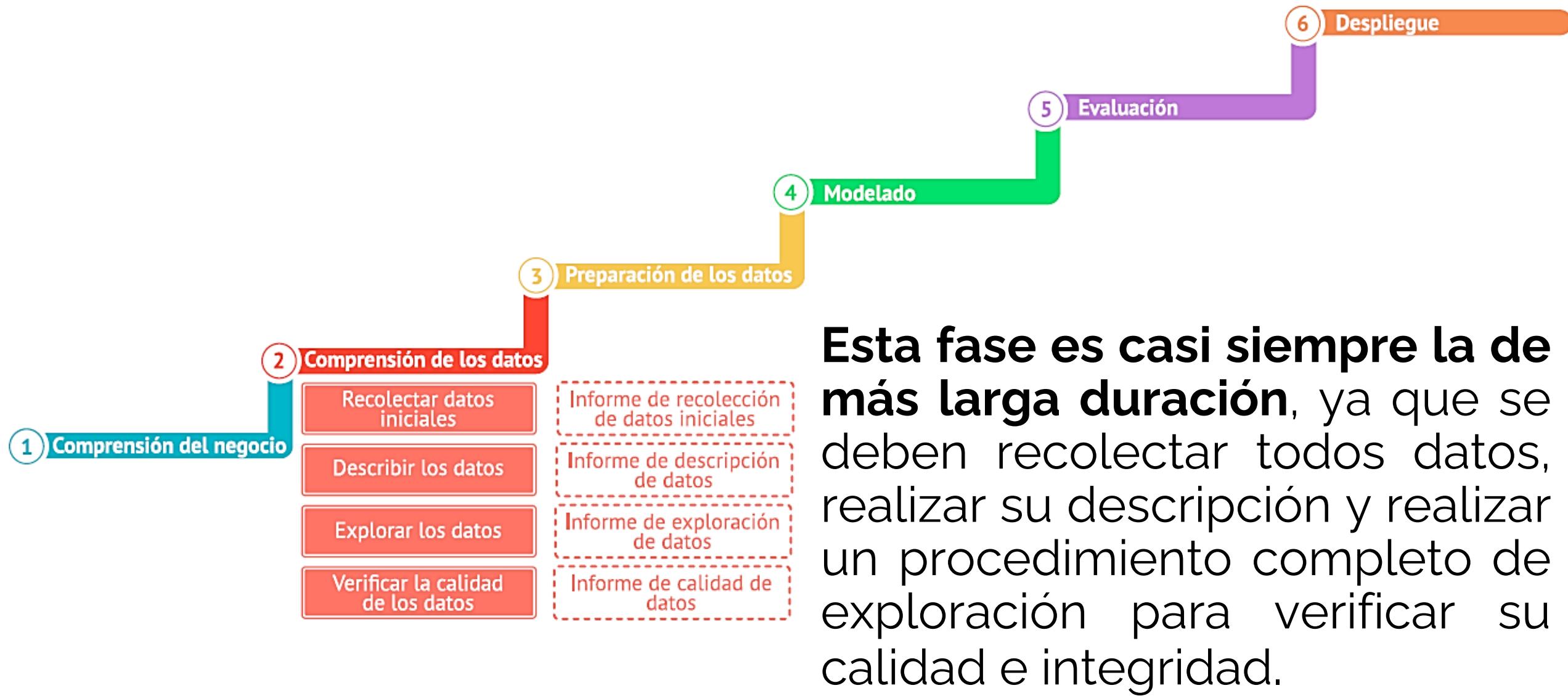
Por ejemplo, en la fase 4 del proceso "Modelado", existe una tarea general llamada "Seleccionar técnica de modelado"; dentro de esta existirán dos tareas especializadas llamadas "Técnicas de modelado" y "Supuestos del modelado". Finalmente, el último nivel define las acciones, decisiones y los resultados sobre el proyecto de Ciencia de Datos.

1. Comprensión del negocio

En esta tarea es necesario obtener toda la información posible sobre los objetivos desde el punto de vista comercial. Esta tarea es fundamental realizarla correctamente, ya que su objetivo es clarificar los problemas que se plantean, definiendo los objetivos y los recursos necesarios.



2. Comprensión de los datos



3. Preparación de los datos

En la fase de preparación de datos, se procederá a **la adaptación del conjunto de datos seleccionados para su utilización en el análisis de datos y Machine Learning**. Se realizarán acciones como seleccionar un subconjunto de datos, limpiarlos para mejorar su calidad o crear nuevos datos a partir de los seleccionados.

6 Despliegue

5 Evaluación

4 Modelado

3 Preparación de los datos

Seleccionar datos

Limpieza de datos

Construir datos

Integrar los datos

Formatear los datos

Inclusión o exclusión de datos
Informe de limpieza de datos

Atributos derivados

Unificación de datos

Informe de formateo de datos

Dataset

Descripción del Dataset

Registros generados

1

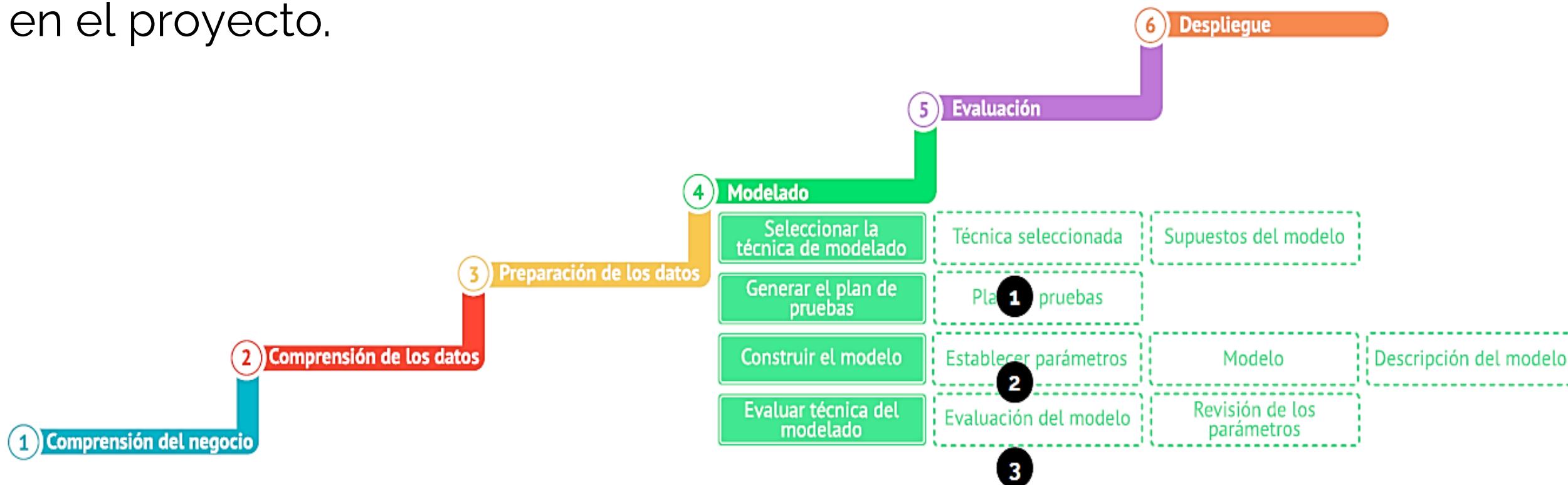
1 Comprensión del negocio

2 Comprensión de los datos

También se estudiarán **las posibles anomalías de los datos**, como huecos vacíos en los atributos y la subsanación en caso de ser motivo de errores o de mantenerlos debido a causas válidas.

4. Modelado

En esta fase, se escogerán las técnicas de Machine Learning (ML) que mejor se adapten a los objetivos (objetivos de Data Science) propuestos en el proyecto.

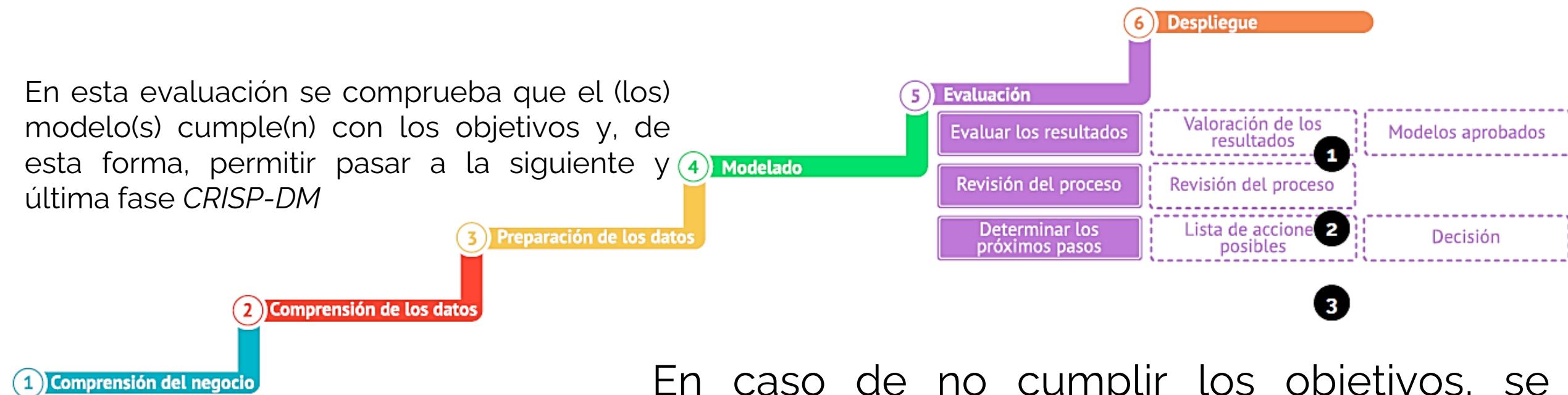


Cada técnica de ML está orientada a resultados diferentes, por lo que unas técnicas son más adecuadas que otras.

5. Evaluación

Esta fase se encarga de **evaluar los modelos generados, pero desde el punto de vista de los objetivos de negocios marcados** en lugar de los objetivos de *Data Science* como en la anterior fase.

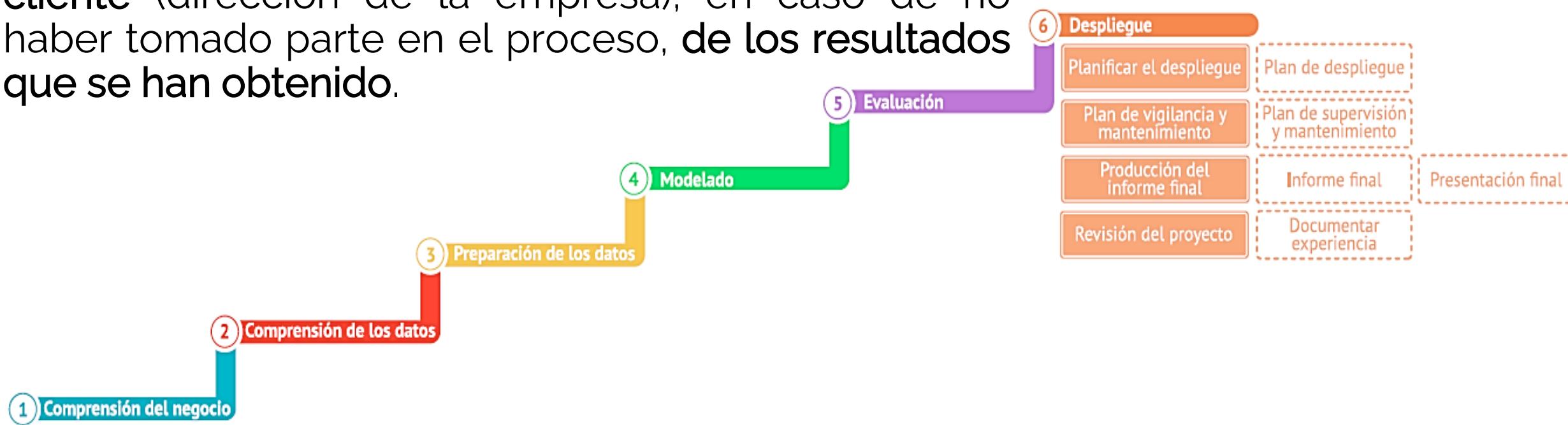
En esta evaluación se comprueba que el (los) modelo(s) cumple(n) con los objetivos y, de esta forma, permitir pasar a la siguiente y última fase *CRISP-DM*



En caso de no cumplir los objetivos, se deberá volver a fases anteriores, donde el problema se haya detectado, tal y como indica el diagrama de fases *CRISP-DM*.

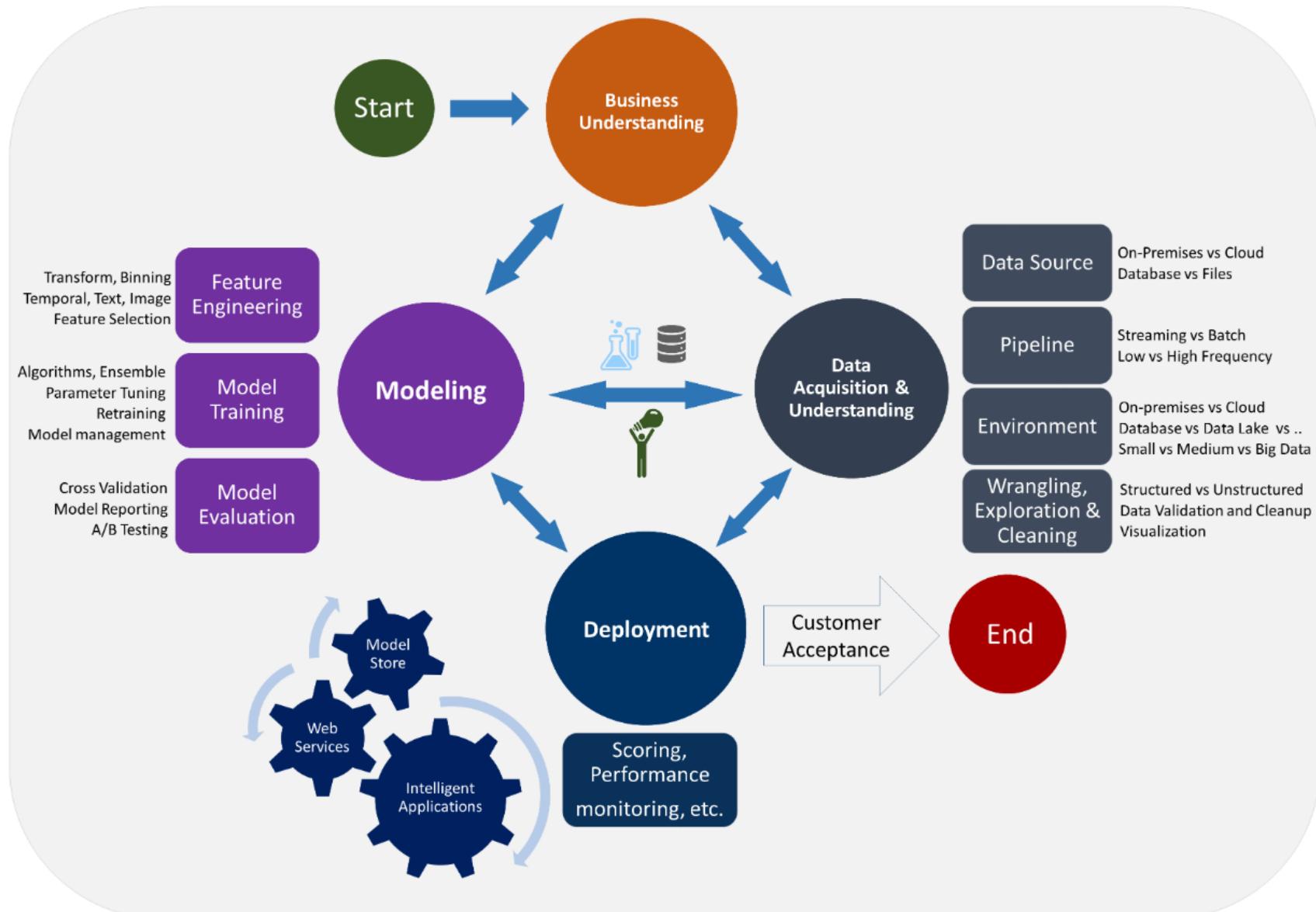
6. Despliegue

Esta es la última fase de la guía *CRISP-DM*, la cual consiste en la puesta en ejecución del proyecto realizado en las fases anteriores, se informará al cliente (dirección de la empresa), en caso de no haber tomado parte en el proceso, de los resultados que se han obtenido.



El ciclo de vida del proceso de ciencia de datos en equipo (TDSP)

Data Science Lifecycle

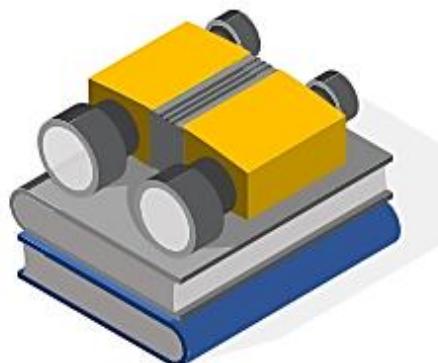


Objetivo final

Diseñar un modelo de operación que permita procesar la data existente y de respuesta a las preguntas de negocio

01 Datos

Identificar las fuentes de información, procesarla y llevarla a un lugar donde se pueda trabajar con ella



02

Información

Reportes iniciales y control de calidad

03

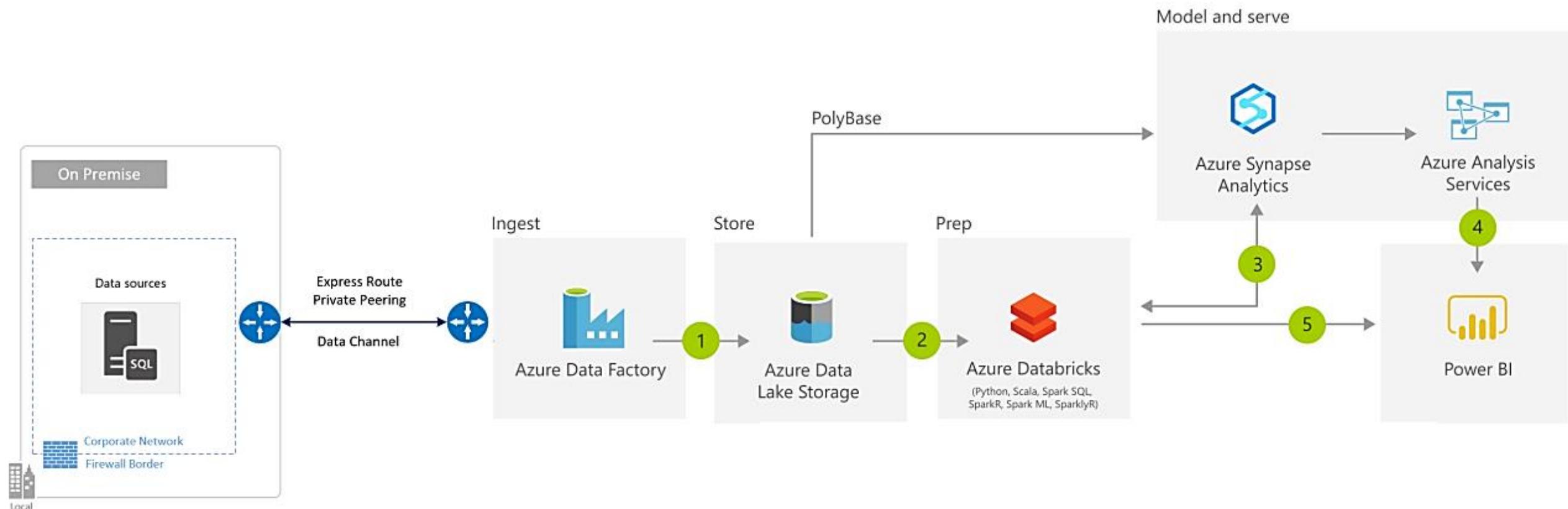
Conocimiento

Construir conclusiones basadas en el análisis de las múltiples fuentes

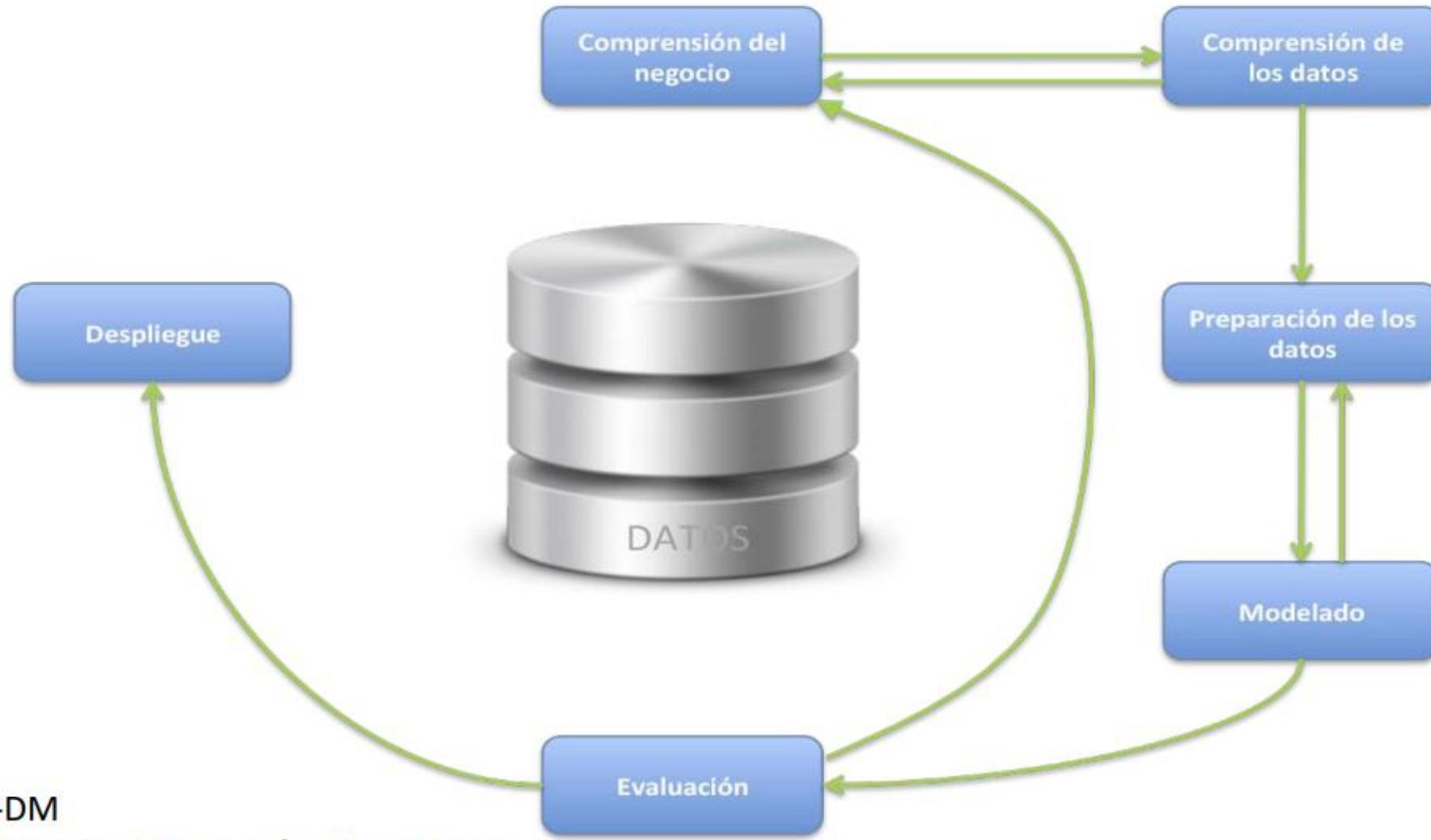




Arquitectura de referencia



Modelo Propuesto



Modelo CRISP-DM

Cross Industry Standard Process for Data Mining

Modelo Propuesto

Modelo CRISP-DM



Entendimiento del negocio

- Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)
- Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio,...)
- Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito)
- Generación del plan del proyecto (plan, herramientas, equipo y técnicas)

Comprensión de datos

- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos
- Verificación de calidad de datos

Preparación de los datos

- Selección de los datos
- Limpieza de datos
- Construcción de datos
- Integración de datos
- Formateo de datos

Modelo Propuesto

Modelo CRISP-DM



Modelado

- Selección de la técnica de modelado
- Diseño de la evaluación
- Construcción del modelo
- Evaluación del modelo

Evaluación

- Evaluación de resultados
- Revisión del proceso
- Establecimiento de los siguientes pasos o acciones

Despliegue

- Planificación de despliegue
- Planificación de la monitorización y del mantenimiento
- Generación de informe final
- Revisión del proyecto



Plan de trabajo

Sprint	Actividad	Entregable/Actividad	Mes 1	Mes 2	Mes 3
1	Entendimiento del negocio	Definición de: <ul style="list-style-type: none">- Requisitos de negocio- Reglas de negocio	<div style="width: 50%; height: 10px; background-color: #90EE90;"></div>		
		- Requisitos funcionales Definición de: <ul style="list-style-type: none">- Requisitos no funcionales- Integraciones externas	<div style="width: 50%; height: 10px; background-color: #90EE90;"></div>		
	Entendimiento de los datos	<ul style="list-style-type: none">- Configuración física de servicio- Mover primera carga a repositorio final- Procesar data- Revisar datos		<div style="width: 100%; height: 10px; background-color: #90EE90;"></div>	



Plan de trabajo

Sprint	Actividad	Entregable/Actividad	Mes 1	Mes 2	Mes 3
2	Preparación de los datos	- Generación datasets finales - Generación de flujos de trabajo de construcción		<div style="width: 100%;"> </div>	
		- Creación de estadística de datos basado en modelos ejecutados		<div style="width: 100%;"> </div>	
	Modelamiento	- Diseño de bodega - Implementación de bodega - Poblado de bodega		<div style="width: 100%;"> </div>	
		- Resultado del cargue		<div style="width: 100%;"> </div>	

Plan de trabajo

Sprint	Actividad	Entregable/Actividad	Mes 1	Mes 2	Mes 3
3	Evaluación	-Integración data final con POWER BI - Creación tablero de control		<div style="width: 100%;"> </div>	
		Ejecución WORKFLOW completo de data y modelos en portal			<div style="width: 100%;"> </div>
	Despliegue	Despliegue en ambiente seleccionado			<div style="width: 100%;"> </div>
		Plan de monitoreo			<div style="width: 100%;"> </div>

Consumo estimado

Microsoft Azure Estimate					
Your Estimate					
Service type	Custom name	Region	Description	Estimated monthly cost	Estimated upfront cost
Storage Accounts		East US 2	Block Blob Storage, General Purpose V2, LRS Redundancy, Hot Access Tier, 500 GB Capacity - Pay as you go, 100,000 Write operations, 100,000 List and Create Container Operations, 100,000 Read operations, 100,000 Archive High Priority Read, 1 Other operations. 1000 GB Data Retrieval, 1000 GB Archive High Priority Retrieval, 1000 GB Data Write	\$10,24	\$0,00
Data Factory		East US 2	Azure Data Factory V2 Type, Data Pipeline Service Type, Azure Integration Runtime: 1 Activity Run(s), 300 Data movement unit(s), 90 Pipeline activities, 90 Pipeline activities – External; Self-hosted Integration Runtime: 1 Activity Run(s), 0 Data movement unit(s), 1000 Pipeline activities, 1000 Pipeline activities – External, 0 x 8 Compute Optimized vCores, 1 x 8 General Purpose vCores, 0 x 8 Memory Optimized vCores, 1 Read/Write operation(s), 1 Monitoring operation(s)	\$698,22	\$0,00
Azure Data Lake Storage Gen1		East US 2	Pay-as-you-go: 16 TB Storage, 0 Read Transactions, 0 Write Transactions	\$638,98	\$0,00
VPN Gateway		East US 2	VPN Gateways, VpnGw3 tier, 1 gateway hour(s), 10 S2S tunnels, 128 P2S tunnels, 0 GB, Inter-VNET VPN gateway type	\$1,25	\$0,00
Azure Databricks		East US 2	Data Analytics Workload, Premium Tier, 1 D3V2 (4 vCPU(s), 14 GB RAM) x 480 Hours, Pay as you go, 0.75 DBU x 360 Hours	\$258,42	\$0,00
Azure Synapse Analytics		East US 2	Tier: Compute Optimized Gen2, Synapse SQL (Provisioned): Compute: DWU 400 x 160 Hours, Storage: 16 TB	\$2,734,08	\$0,00
Azure Analysis Services		East US 2	Standard S1 (Hours), 1 Instance(s), 320 Hours	\$649,60	\$0,00
Power BI Embedded		East US 2	1 node(s) x 200 Hours, Node type: A1, 1 Virtual Core(s), 3GB RAM, 1-300 Peak renders/hour	\$201,62	\$0,00
Support			Support	\$0,00	\$0,00
			Licensing Program	Microsoft Online Services Agreement	
			Total	\$5,192,41	\$0,00



Próximos pasos

Negocio

1. Decidir si se acepta el demo.
2. En caso afirmativo, Programar reunión de entendimiento del negocio.

Tecnología

1. Informar a Microsoft de la decisión.
2. Preparar ambientes.

Modelo

1. Revisar el modelo que mas se ajuste a la necesidad.

Actividad 3 (No hacer)



Preparar una presentación con diapositivas en PowerPoint (u otra herramienta), donde se expliquen las tareas principales y subtareas de cada fase de CRISP-DM consultando los siguientes documentos:



https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPD.pdf

<https://e-archivo.uc3m.es/rest/api/core/bitstreams/714c5452-962e-44cf-993f-ebb3088d4aa5/content>
(páginas 21-33)



Considerar también la estética, imaginar que es para un público no técnico.
Se pueden usar otras referencias si así lo prefiere.