DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Machine Learning Operations (MLOps) Clase 2

Leticia Rodríguez

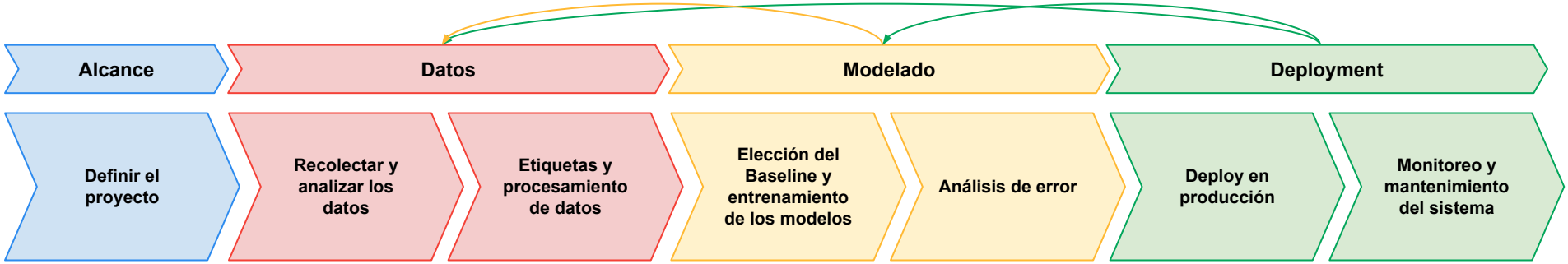
Septiembre 2024 - 2do Cuatrimestre - 4to. Bimestre

Universidad de Buenos Aires - FCEyN - Departamento de Computación

Encuesta y Asistencia

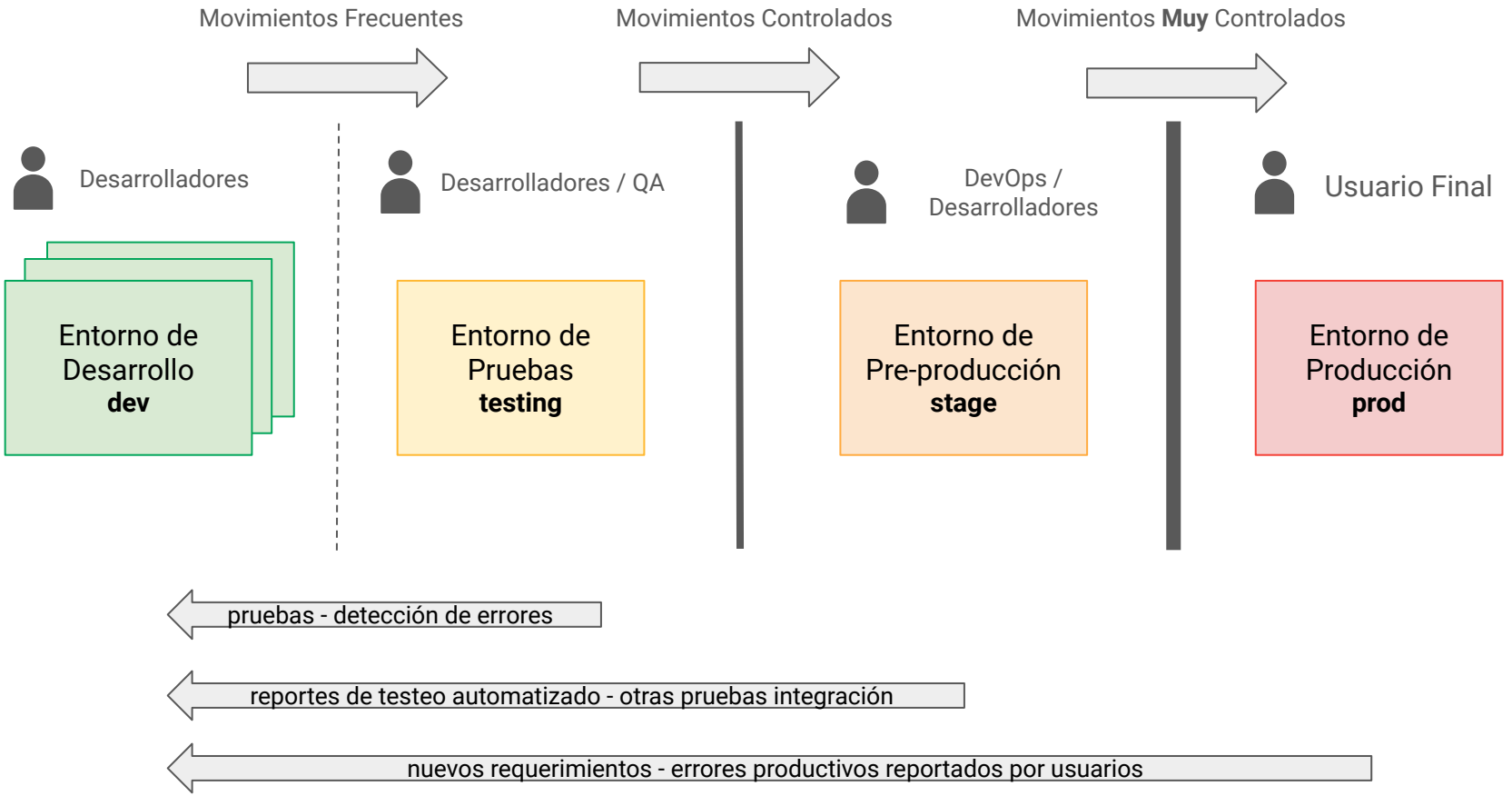
Kahoot de respaso

El flujo de creación de un modelo de ML



Importante: las diferentes etapas tienen flechas que hace que se itere entre las etapas anteriores

Los estándares: Ambientes del desarrollo de Software



Actividad: Machine Learning Operations (MLOps)

En grupos analicen el paper, Machine Learning Operations (MLOps): Overview, Definition, and Architecture. 2022.

1. Lean el abstract, las conclusiones y el primer párrafo de la Introducción y expliquen: ¿por qué es necesario el MLOps?
2. Lean la sección: Foundation of DevOps. Discutan en el grupo que entienden por DevOps a partir de su propia experiencia y la lectura del paper.
3. Lean los principios que describen los autores sobre MLOps en la sección 4.1 Principles.
4. Comparen los roles definidos en la sección 4.3 con los visto la clase pasada
5. Discutan la figura 4: End-to-end MLOps architecture and workflow with functional components and roles
6. Finalmente, lean la sección 6. Conceptualización
7. Por último, hagamos una puesta en común en el curso

Ref: https://arxiv.org/abs/2205.02302

MLOps

MLOps (Machine Learning Operations) es un paradigma que incluye aspectos como las mejores prácticas, conjuntos de conceptos y una cultura de desarrollo en lo que respecta a la conceptualización, implementación, monitoreo, despliegue y escalabilidad de extremo a extremo de productos de aprendizaje automático. Sobre todo, es una práctica de ingeniería que aprovecha tres disciplinas contribuyentes: aprendizaje automático, ingeniería de software (especialmente DevOps) e ingeniería de datos. MLOps tiene como objetivo producir sistemas de aprendizaje automático al cerrar la brecha entre el desarrollo (Dev) y las operaciones (Ops). Básicamente, MLOps tiene como objetivo facilitar la creación de productos de aprendizaje automático al aprovechar estos principios: automatización de CI/CD, orquestación de flujo de trabajo, reproducibilidad; control de versiones de datos, modelos y códigos; colaboración; capacitación y evaluación continuas de ML; seguimiento y registro de metadatos de ML; monitoreo continuo; y bucles de retroalimentación.
De la publicación
https://arxiv.org/pdf/2205.02302

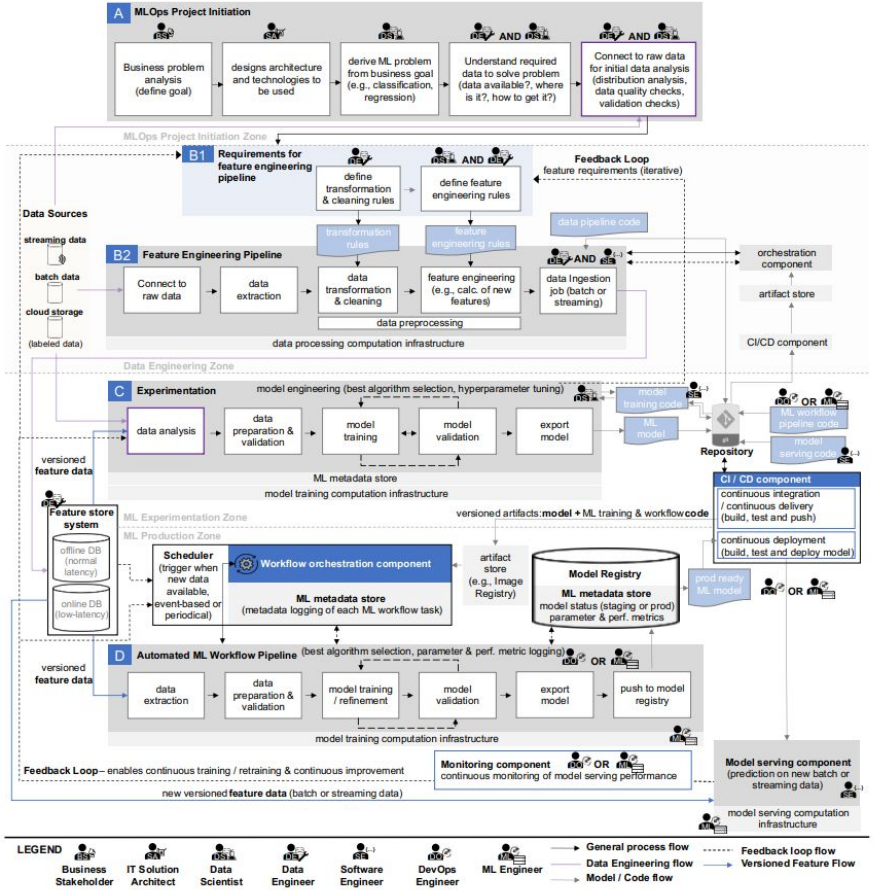
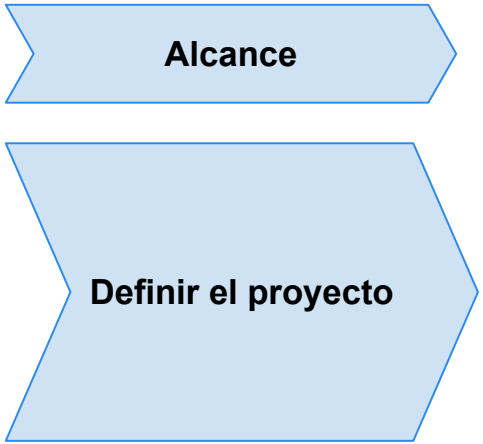


Figure 4. End-to-end MLOps architecture and workflow with functional components and roles

Definición el proyecto de Machine Learning



Actividad: Creemos un recomendador de productos

Imaginemos que estamos trabajando en una compañia de comercio electrónico y se nos pide, como equipo de ciencia de datos, un recomendador de productos.

El CEO quedó fascinado con las recomendaciones de una páginas de películas. Se acercó al Product Manager y le pidió hacer algo similar en su compañía de comercio electrónico de electródomésticos:

1. ¿Qué preguntas le harian a los stakeholders sobre lo que quieren tener?
2. ¿Qué precisarían o como organizarían el trabajo alredor del requerimiento?
3. ¿Cuántas personas precisarían para el trabajo, tiempo y descripción de los recursos?
4. ¿Qué otras consideraciones deberían tener en cuenta?

Requerimientos para Sistemas de ML

- **Confiable** (Reliability): El sistema debe funcionar correctamente. El sistema no debe fallar y estar disponible. En Ingeniería del Software se pueden detectar errores o problemas de infraestructura que tiene el softwarea travéz de los logs o monitoreo. Dada la naturaliza del Machine Learning, un sistema que empieza a dar predicciones incorrectas es también un punto de fallo.
- **Escalable** (Scalability): Los sistemas pueden crecer en tráfico, cantidad de datos, en cantidad de modelos, en infraestructura y incluso en funcionalidades, nuevos requerimientos llegan día a día. El sistema tiene que tener la capacidad de soportar estos cambios sin afectar a los usuarios.
- **Mantenible** (Mainteinability): Tiene que haber, en la empresa o en el laboratorio, personal que pueda lidear con el día a día de los sistemas de ML y su integración con el resto del software. Esto significa garantizando el correcto funcionamiento y actualizando las versiones o nuevos modelos.
- **Adaptable** (Adaptability): Un tema importante es la producción de cambios en los datos o en el concepto propio que fundamentan el modelo. Esto se conoce como drifts. El sistema tiene que anticiparse al que ocurra y tener la posibilidad de adaptar el sistema a los cambios de esta u otra naturaleza.

IA Responsable

- Hace tiempo, los investigadores vienen debatiendo sobre la Éticas de la AI, lo que llamaron últimamente como **Inteligencia Artificial Responsable**.
- En esta intersección entre la Ingeniería y el AI, muchas veces se pierde dimensión del cuándo, cómo y dónde debemos usar estos algoritmos de AI siendo **responsables sobre el impacto social y humano**
- Tradicionalmente se definen 4 dimensiones en las cuales se puede evaluar estos sistemas de AI:
  - **Equidad (Fairness)**
  - **Responsabilidad**
  - **Seguridad**
  - **Privacidad**
- Es necesario durante la definición del proyecto de Machine Learning, **hay que evaluar la propuesta y diseño del sistema en AI Responsable.** Plantearse posibles escenarios donde dichos supuestos pueden no cumplirse y realizar **mitigaciones**, es decir, **acciones que lleven a anular o mitigar daños.**

Riesgos y Prueba de Concepto

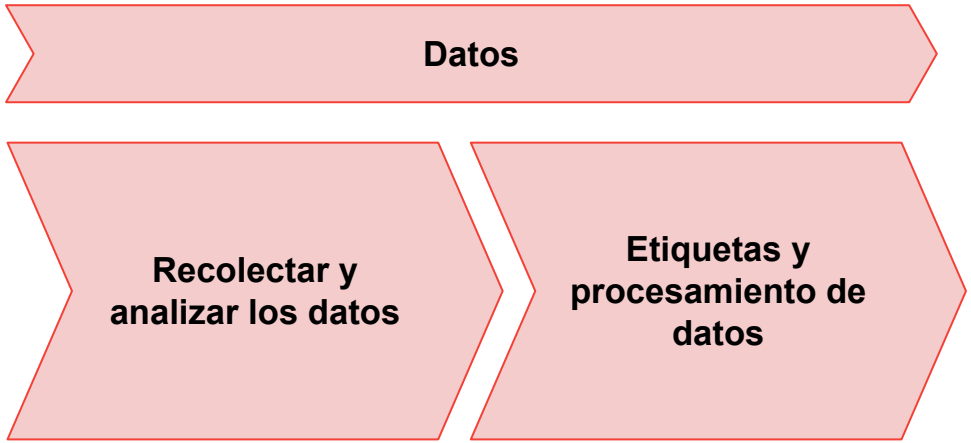
- **Primero: Definición del problema a resolver**
- **Con respecto a la funcionalidad en el sistema:**
    - Nueva funcionalidad
    - Funcionalidad existente que no usa ML
    - Funcionalidad existente y ya usa ML
- **Esto conlleve diferentes desafíos:**
    - **Funcionalidad existe:**
        - ¿Por qué queremos un nuevo modelo?
        - ¿Qué metricas buscamos mejorar?
        - ¿Cómo vamos a comparar el modelo nuevo con el viejo?
        - ¿Cómo vamos a integrar el modelo nuevo y deprecar el viejo?
        - El impacto que podria tener el cambio en los usuarios finales
    - **Funcionalidad no existe:**
        - ¿Cómo garantizar que el modelo nuevo performa como esperamos?
        - ¿Dónde ubicaríamos el nuevo modelo y que otros desarrollo de software son necesarios para hacer que el modelo llegue al usuario?
        - ¿Qué metricas de negocio impactaría el nuevo desarrollo (a nivel sistema)?
    - ¿Podemos hacer **pruebas de concepto** para tomar la decisión de negocio?

Break
15 minutos
y seguimos con MLOps

Prueba de Concepto en Machine Learning

- Pequeña prueba que busca demostar el funcionamientoy factibilidad de un sistema
- En la construcción de Sistemas de ML, significa demostrar que se puede armar un modelo o un sistema de ML para resolver a un problema
- La PoC puede incluir alguna especie de interfaz de usuario o API pero no necesariamente. La idea es demostrar cómo usando ML se puede resolver un problema y decidir a partir de ahí el tiempo, recursos e inversión requerida para hacer la funcionalidad completa (para desarrollar un sistema de ML completo que sea productivo y este disponible para ser usuado por los usuarios)
- Mucha veces, la PoC incluyen desarrollar algun pequeño modelo o usar API que llamen a modelos pre-entrenados pero con una precisión menor o sobre un conjunto acotado de datos
- El objetivo final es mostrar la idea a las personas que deciden sobre los proyectos o la dirección del negocio

Procesamiento de Datos



La importancia de los Datos: Mind vs. Data

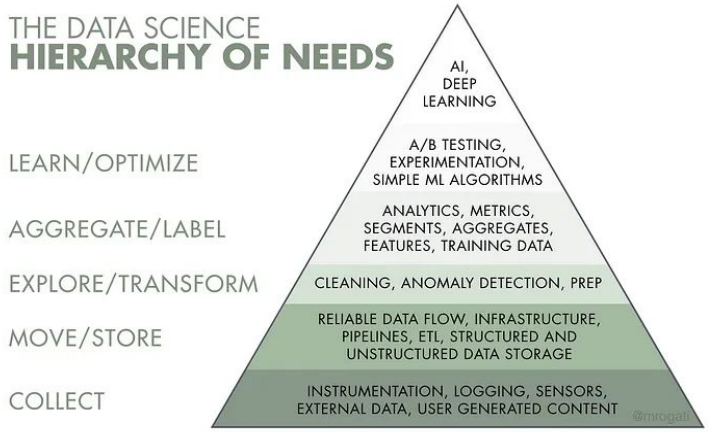
¿Qué opinan?

¿Es más importante trabajar en mejorar los algoritmos de los modelos o incrementar la cantidad de datos?

¿Cómo son los modelos actuales en relación a los datos y capacidades de los modelos de aprender sobre pocos datos?

La importancia de los Datos: Mind vs. Data

- Sigue siendo un debate continuo en donde focalizar los esfuerzos de la comunidad científica respecto a la creación de nuevos algoritmos que utilicen menos datos.
- Areas de investigación como meta-learning, algoritmos zero-shot, few-shot se ocupan de esto.
- Actualmente, con el surgimiento de las LLMs y el impacto que estas parecen tener en la sociedad, se refuerza la teoría que, a día de hoy, muchos modelos de Machine Learning están fuertemente anclados en los datos, por los cual, cantidad y calidad de los datos son fundamentales para sacar el máximo provecho de las técnicas actuales.
- Estos debates atraviesan la comunidad científica.



Credit: Mónica Rogatti - “The AI Hierarchy of Needs”

Recolección y almacenamiento de datos

- Necesidades de datos
  - Entrenamiento del modelo, testeo del modelo y predicción/uso del modelos
- Consideraciones
  - Tipo de modelo a construir
  - Cantidad de datos
  - Pre-procesamiento
  - Formato de los datos y almacenamiento
  - Selección de features
  - Etiquetado
  - Metadata y trackeo de los datos usados para entrenamiento y testeo
  - Posibles necesidades de incrementar la cantidad de datos
  - Diferencias de origen entre datos de entrenamiento, testeo y predicción
    - Por ejemplo, en caso que predicciones online estas podrían venir de data dinámica y generada por usuarios.
  - Costos - almacenamiento, equipos, personal

Fuentes de Datos

- Estructurada
  - Base de datos relacionales: PosgreSQL, MySQL, DBs Cloud
  - NoSQL
    - Archivos de texto con formato: json, xml, csv
    - GraphSQL
- No estructurada
  - Imágenes
  - Archivos pdf
  - Audio
  - Video
- Muchas veces los datos no estructurados pueden estar acompañados por meta-datos: versionado, fechas, ubicación, tags.

Fuentes de Datos

Base de datos SQL

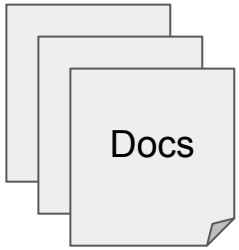
Relacional

id cliente	Nombre	Apellido
1	Jose	Pepe
2	Hernan	Perez
3	Ana	Rodriguez
4	Julia	Rodriguez
5	Mar	

id venta	id product	cantidad
1	1001	10
2	1001	1
3	1002	1
4	1008	1

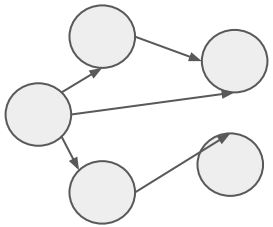
Oracle, AlloyDB, MySQL, PosgreSQL, DB2

Base de datos NoSQL



Documental

MongoDB, DynamoDB



Grafos

Neo4j

Clave 1	Valor 1
Clave 2	Valor 2
Clave 3	Valor 3

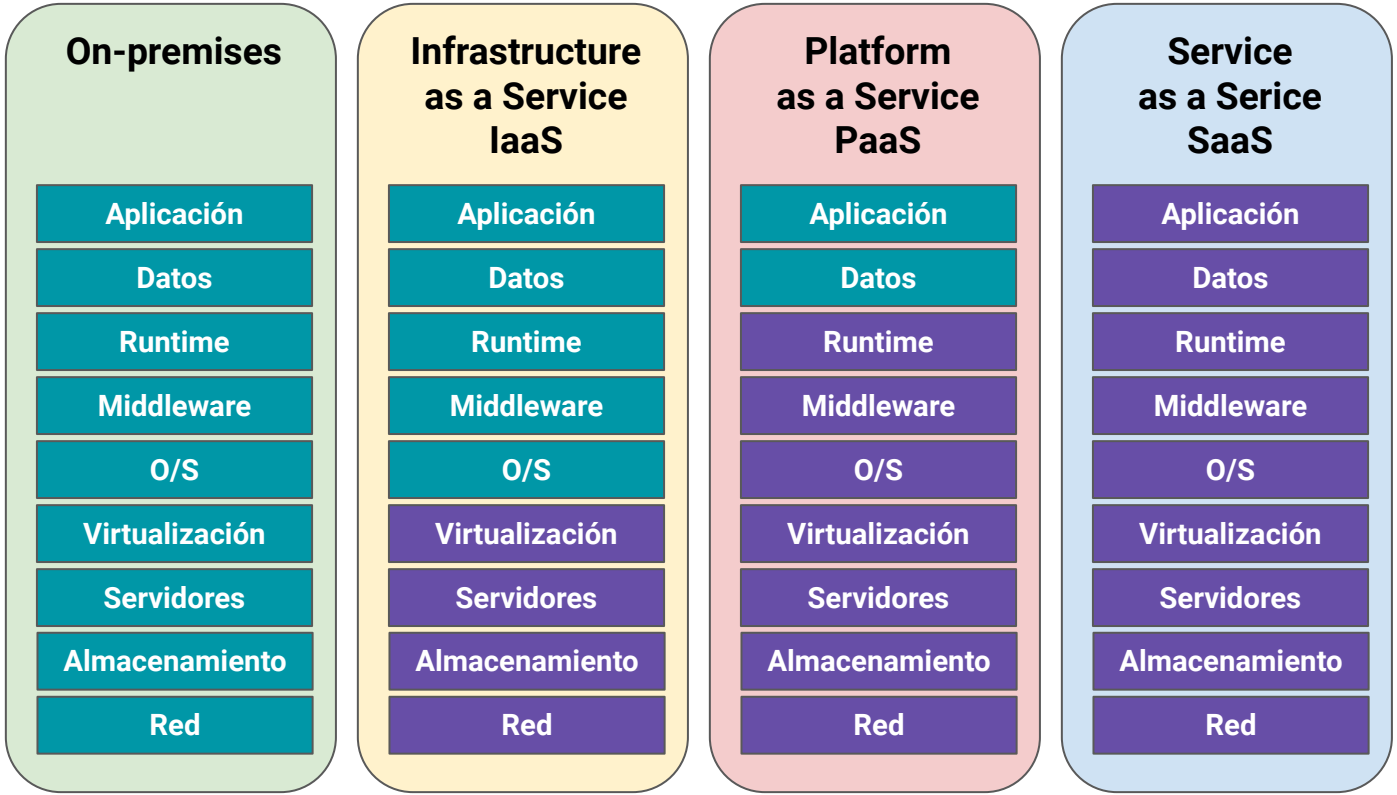
Clave Valor

Redis - Cassandra

¿Cómo almacenar y compatir millones de datos?

- **Infraestructura contratada propietaria**: donde la empresa/laboratorio hace el mantenimientos de los equipos y del software específico necesario: por ejemplo, tenes un servidor en la red con bases de datos MySQL con distintos accesos y permisos.
- **La nube**: distintos proveedores de cloud que se ocupan del manteniemiento y seguridad de los datos. Estos proveen diferentes servicios para el almacenamiento y trackeo de data estructurada y no estructurada. También ofrecen integraciones con diferentes almacenamientos de datos: bases de datos de renombre, NoSQL, etc. Las nubes incluso ofrecen servicios de baja latencia para grandes datos.
- Esto refiere a dos **modelos de negocios CapEx y OpEx**. Estos modelos hablan sobre la inversión en capital de la infraestructura vs. el costo operacional de usar los servicios de la nube - modelo “pago por uso”.

On-premises, SaaS, PaaS, IaaS en la nube



Ejemplos - IaaS

Instance Type	CPU type	CPU GHz/ RAM/SSD	Price $/Month
Basic-2	Intel	2.3/4/80	24.00
Premium-2	Intel Cascade L	2.5/4/80	28.00
Premium-2-AMD	AMD Rome	2.0/4/80	28.00
CPU-opt-2 (S)	Intel Skylake	2.7/4/25	42.00
CPU-opt-2 (C)	Intel Cascade L	2.7/4/25	42.00
CPU-opt-2 (I)	Intel Ice Lake	2.6/4/25	42.00

What is Iaas? - RedHat

Ejemplos - SaaS

- Servicios de streaming de películas , videos
- Redes sociales
- Servicios de mail
- Servicios de documentos, planillas de cálculo, suites de oficina

Costos y disponibilidad de los datos

Los datos en la nube pueden estar almacenados en diferentes tipos de almacenamiento con costos relativos a la frecuencia de acceso a los datos.

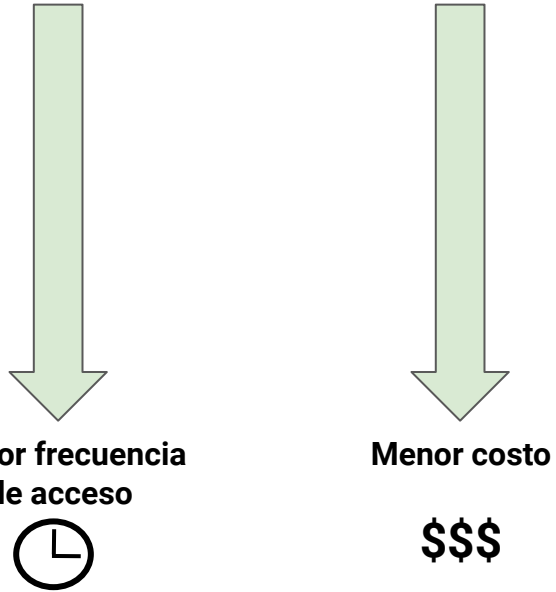
Definimos que los almacenamientos pueden ser:

Clásico

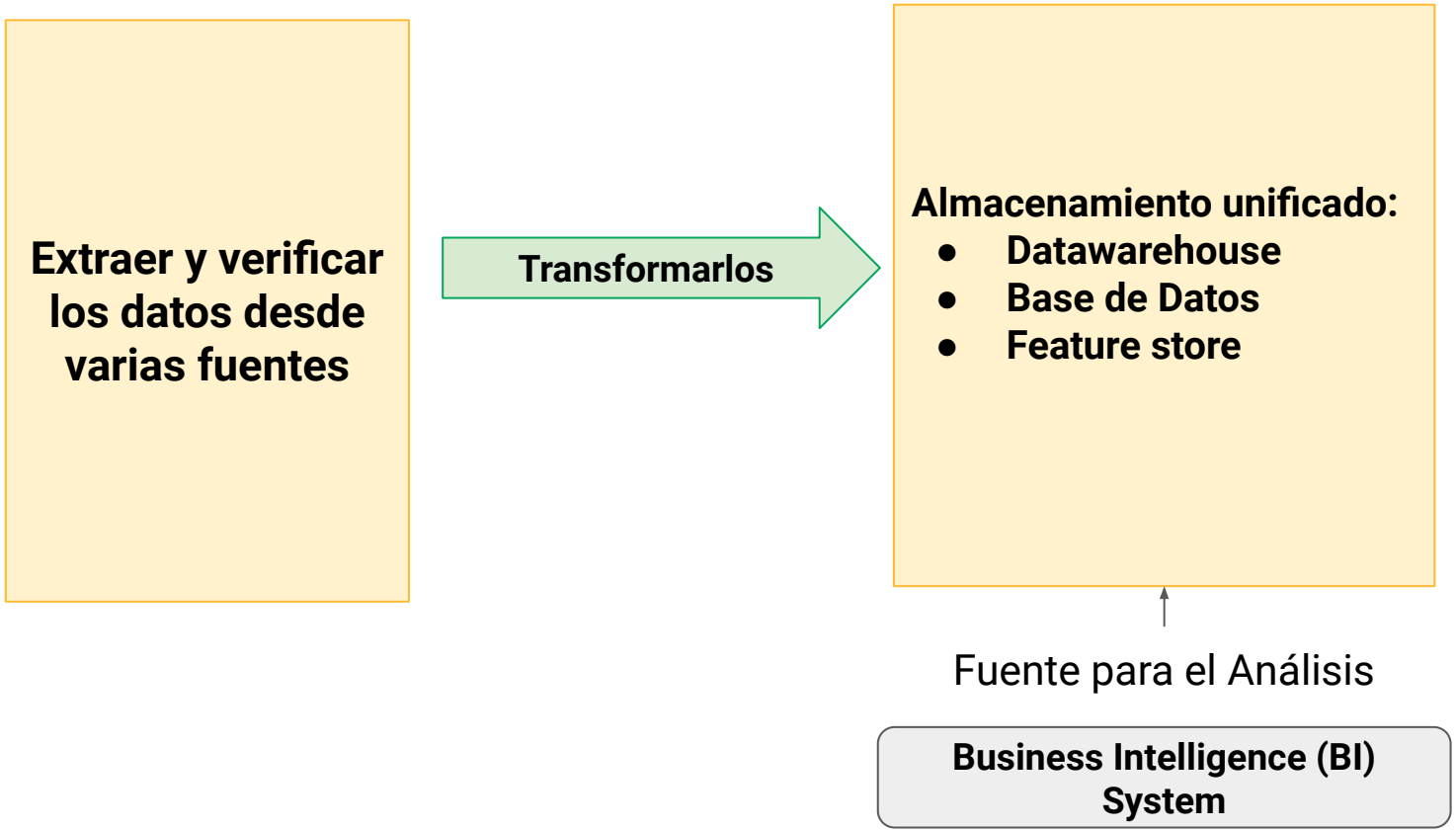
Nearline

Coldline

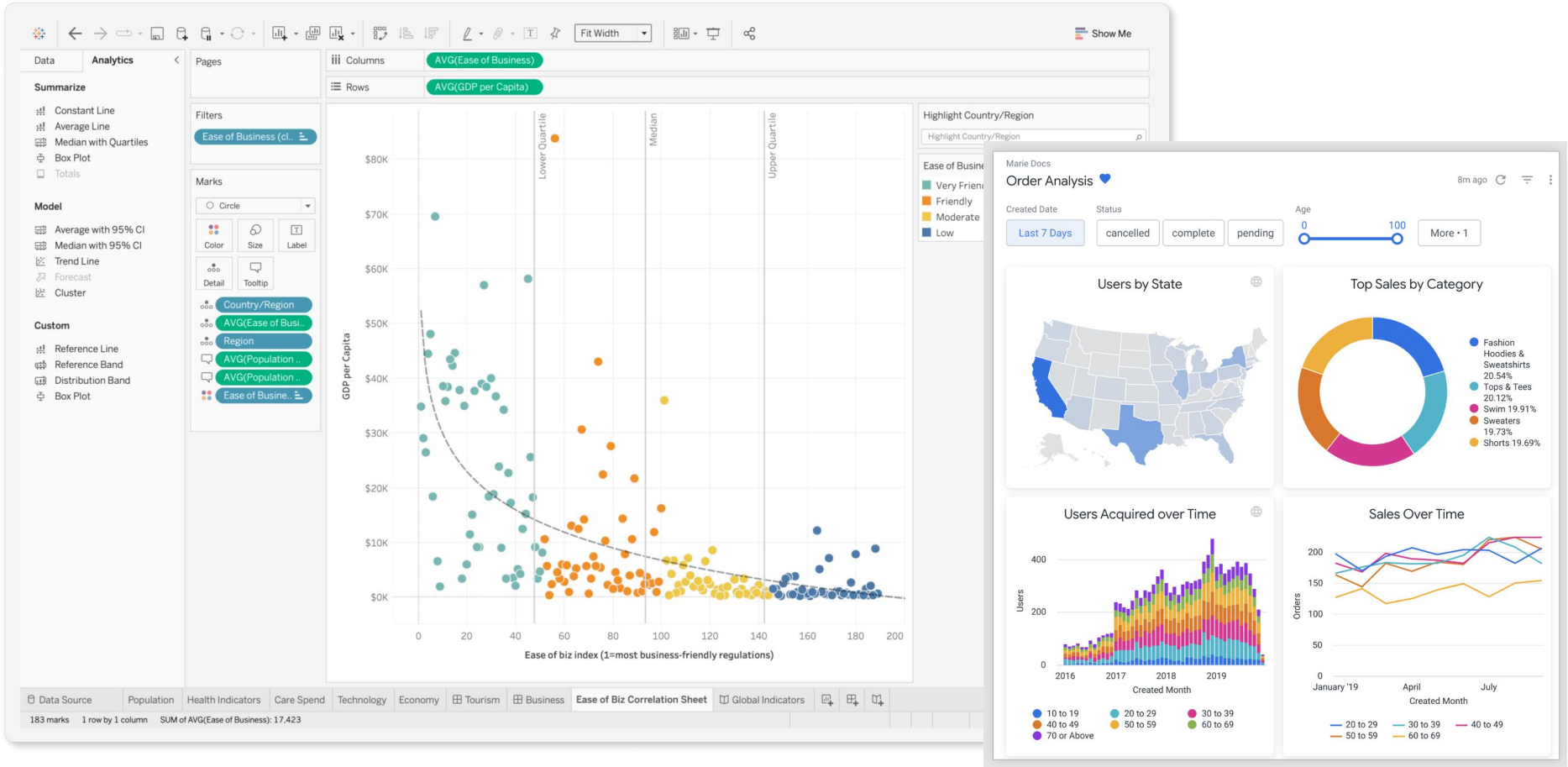
Archivo



ETL - Extract, Transform, Load



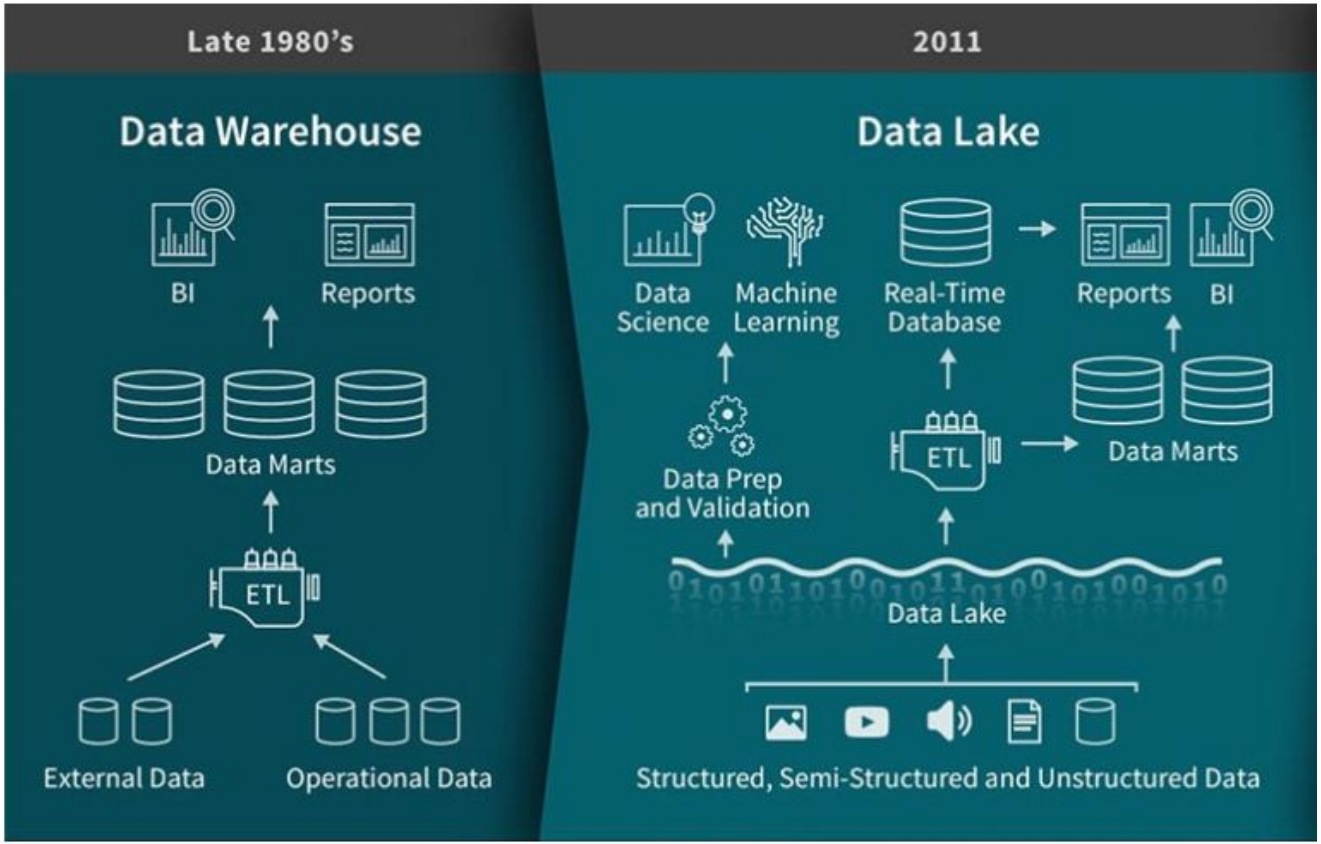
Business Intelligence - BI



Data Lakes

- Los Data Lakes proveen un único almacenamiento de un alto volumen datos crudos, estructurados y no estructurados.
- El objetivo es:
  - Reducir el coste total de propiedad.
  - Simplificar la gestión de los datos.
  - Preparar la incorporación de inteligencia artificial y aprendizaje automático.
  - Agilizar las analíticas.
  - Mejorar la seguridad y el gobierno.
- Constituyen una fuente de datos única para proyectos de Análisis de Datos y Aprendizaje Automático

Data Lake


Late 1980's
2011
Data Warehouse
BI
Reports
Data Marts
ETL
External Data
Operational Data
Data Lake
Data Science
Machine Learning
Real-Time Database
Reports
BI
Data Prep and Validation
ETL
Data Marts
Data Lake
Structured, Semi-Structured and Unstructured Data

Estrategias de sampleo de datos

- Los datos en producción pueden ser muchos más de los necesarios para crear el modelo.
- Pueden encontrarse distribuídos por toda la empresa y hasta replicados
- La elección de los datos para el armado de los datasets de entrenamiento, validación y testeo puede ser relizada con distintas técnicas de sampleo.
- Técnicas de sampleo:
  - Nonprobility Sampling: Convenience, Snowball, Judgement, Quota
  - Simple Random Sampling
  - Stratified Sampling
  - Weighted Sampling
  - Reservoir Sampling

EDA - Exploratory Data Análisis

- En simultáneo o luego de la obtención de datos, todo proyecto de Machine Learning arranca por el análisis exploratorio de los datos
- Es decir, buscamos entender la naturaleza de los datos respondiendo preguntas cómo:
  - Cantidad de clases o variables objeto de predicción
  - Correlación entre características (features)
  - Features candidatas que pueden ser relavantes para la predicción
  - Valores máximos, mínimos, medianas, media, modo
  - Detección de variables no numéricas, categóricas
  - Detección de Outliers
  - Posible preprocesamiento de tokens y vocabulario necesario para NLP. Análisis de texto.
  - Características de los sets de imágenes, tamaños en píxeles, variabilidad
  - Detección de valores nulos
  - Cantidad de Datos
  - Distribución - Desbalanceo

Pre-procesamiento de datos - Features Engineering

- Los datos son imperfectos. Los modelos de Machine Learning pueden requerir que los datos estén organizados de determinada manera para facilitar el aprendizaje del algoritmo.
- El objetivo del pre-procesamiento de datos funciona en dos etapas: predicción y entrenamiento. Puede variar según el algoritmo de ML a usar.
- Dependiendo de los datos y el algoritmo a utilizar precisaremos procesos:
  - para la eliminación de outliers
  - para la normalización de features
  - compleción o eliminación de datos incompletos (imputation methods)
  - de combinación de features, de ser necesario
  - de conversión de features, one-hot encoding, categoricas, escalado, estandarización, transformaciones log, discretización
  - de pre-procesamiento de imágenes: escalado, conversión a grises
  - de pre-procesamiento de texto: tokenización, eliminación de stopwords, armado de vocabulario, encodeo
  - creación de embeddings

Privacidad - Información Personal

- Hay consideraciones importantes en cuento a la Privacidad y Seguridad de nuestros valores éticos en Inteligencia Artificial:
  - Propiedad de los datos - Es importante tener en cuenta quien es el propietario de los datos, ya sea del dataset como de la información personal confiada a las empresas o laboratorio.
  - Manipulación de los datos - Trabajar con datos personales o propietarios conlleva una responsabilidad. Así los datos de producción, son cuidados y su acceso es restrigido, contar con dichos accesos no significa estar habilidados a sacarlos del ambiente o distribuirlos. En cualquier caso, tener en cuenta situaciones en que los datos deben ser preprocesados para eliminar información:
    - personal (PII - Personal Identifiable Information)
    - identificable médica (PHI - Personal Health IInformation)

Actividad: Desafíos Éticos en el trabajo con Datos

Como actividad, revisemos el código de Ética de la ACM: https://www.acm.org/code-of-ethics. Este código es usado por algunas conferencias científicas para hacer sus revisiones éticas.
Vamos a leer las secciones 1 y 2.

- En grupo:
  - Lean 3 subsecciones elegidas al azar dentro de esas secciones
  - Compartan lo que entienden del texto y relacionénlo con lo visto hasta ahora en la materia.
  - Comenten experiencias propias, opiniones y criterios con respecto al texto.
- Hagamos una puesta en común sobre lo más interesante que ha surgido en los grupos.

Material Recomendado de esta semana

Andrew Ng: Bridging AI's Proof-of-Concept to Production Gap (2020)

Chip Huyen, Designing Machine Learning Systems, 2022 - Chapter 4: Training Data - Sampling - Pág. 82-87