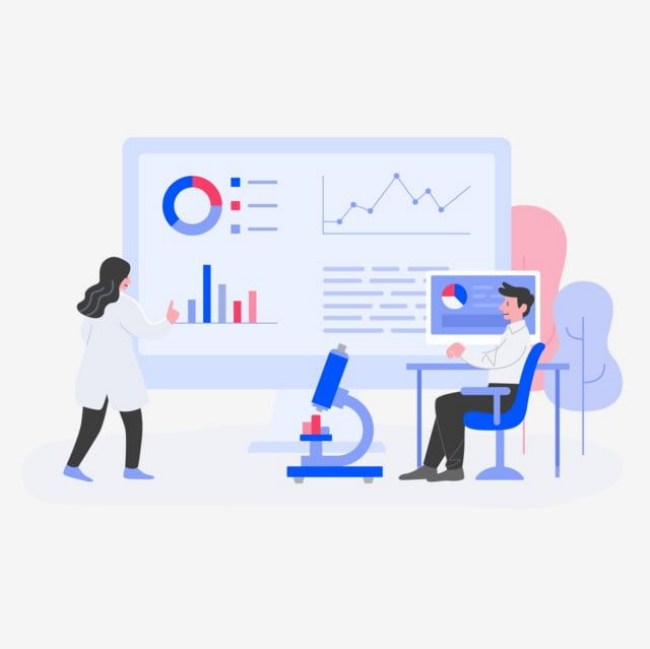
Escuela de Datos Vivos

Curso Data Engineer: Creando el pipeline de datos

MÓDULO A-1

¿Qué hace un ingeniero de datos?

Diseño, construcción, mantenimiento de los pipelines de datos.



Trabajar con big data

- Manejo de diferentes tipos de fuentes (archivos, DBS, aplicaciones, IOT, etc.)
- Grandes volúmenes de información
- Creación de tablas/vistas para contestar consultas típicas del negocio
- Sistema de computación distribuida
- Buenas prácticas

Ciclo de vida

Ciclos de vida del dato

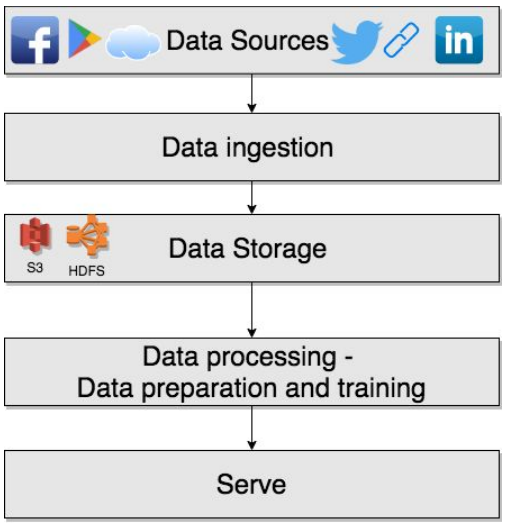
- **Ingesta:** Aplicaciones, Streaming data, Batch data
- **Almacenamiento:** patrones de acceso a la información, control de acceso, Tiempo de almacenamiento
- **Proceso & Análisis:** data transformations, data analysis
- **Visualización**

Ciclo de vida

Ingesta

Ingesta

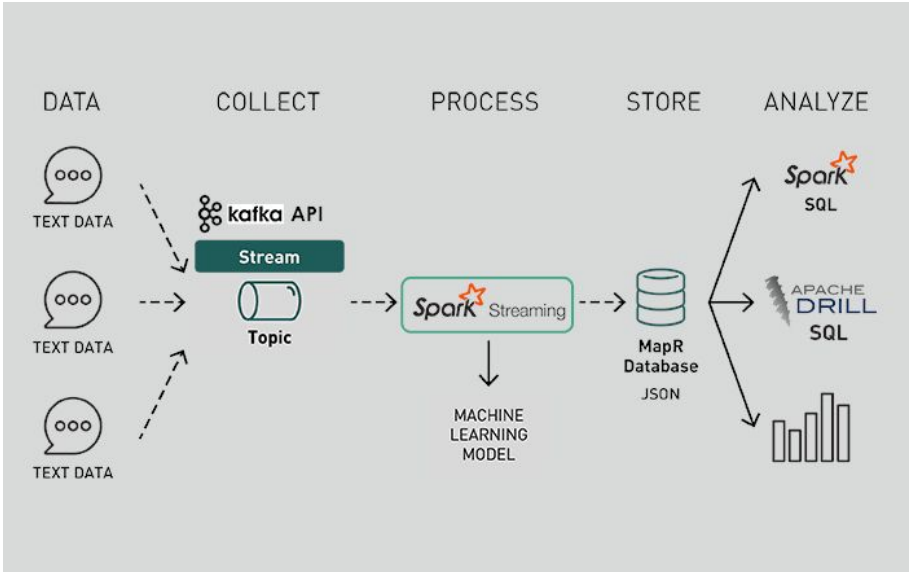
Aplicaciones



fuente: towardsdatascience.com

Ingesta

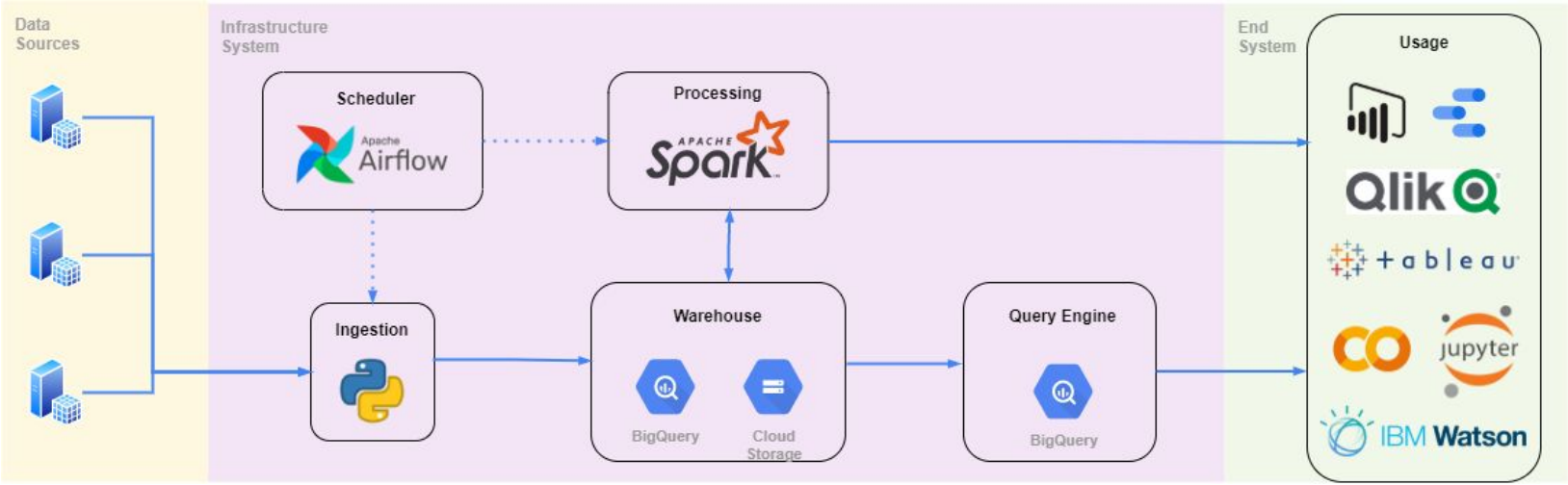
Streaming data



fuente: developer.hpe.com

Ingesta

Batch data



fuente: medium.com

Ingesta

- Aplicaciones
- Streaming data
- Batch data

Ciclo de vida

Almacenamiento

Almacenamiento

Patrones de acceso a la información

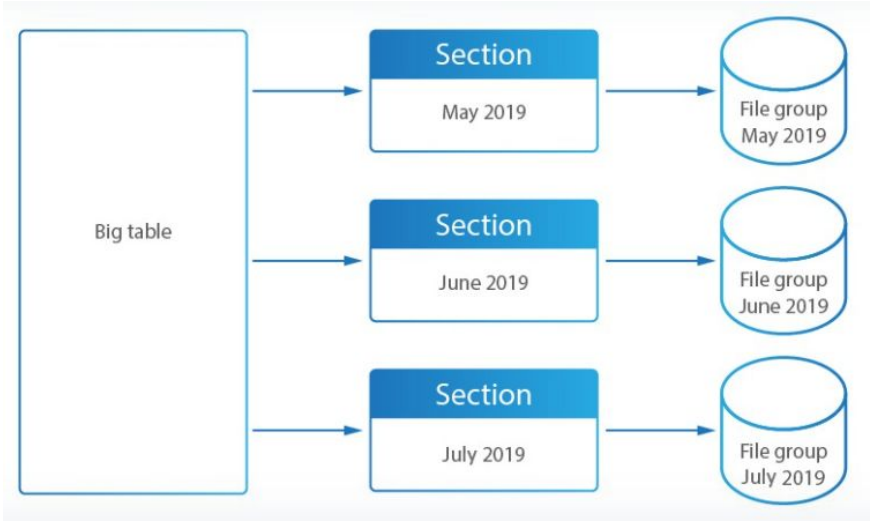
Access pattern	Priority	Read or write	Description	Type (single item / multiple items / all)	Key attribute	Filters	Result ordering
Create user profile	*High*	*Write*	*User creates a new profile.*	*Single item*	*Username*	*N/A*	*N/A*
Update user profile	*Medium*	*Write*	*User updates their profile.*	*Single item*	*Username*	*Username = current user*	*N/A*
Get user profile	*High*	*Read*	*User reviews their profile.*	*Single item*	*Username*	*Username = current user*	*N/A*
Create a game	*High*	*Write*	*User creates a new game.*	*Single item*	*GameID*	*N/A*	*N/A*
Find open games	*High*	*Read*	*User searches for open games. Search results are sorted by start timestamp in descending order.*	*Multiple items*		*GameStatus = open*	*Start timestamp descendent*

fuente: amazon.com

- **Tamaño**
  - partición

Almacenamiento

Patrones de acceso a la información



- **Tamaño**
  - partición

Esquema de organización de datos en el que los datos de la tabla se dividen entre varios almacenes.

fuente: amazon.com

Almacenamiento

Patrones de acceso a la información

Access pattern	Priority	Read or write	Description	Type (single item / multiple items / all)	Key attribute	Filters	Result ordering
Create user profile	*High*	*Write*	*User creates a new profile.*	*Single item*	*Username*	*N/A*	*N/A*
Update user profile	*Medium*	*Write*	*User updates their profile.*	*Single item*	*Username*	*Username = current user*	*N/A*
Get user profile	*High*	*Read*	*User reviews their profile.*	*Single item*	*Username*	*Username = current user*	*N/A*
Create a game	*High*	*Write*	*User creates a new game.*	*Single item*	*GameID*	*N/A*	*N/A*
Find open games	*High*	*Read*	*User searches for open games. Search results are sorted by start timestamp in descending order.*	*Multiple items*		*GameStatus = open*	*Start timestamp descendent*

fuente: amazon.com

- **Tamaño**
  - partición
- **Forma**
  - velocidad
  - escalabilidad
- **Velocidad**
  - particiones
- **Prioridad**

Almacenamiento

Control de acceso

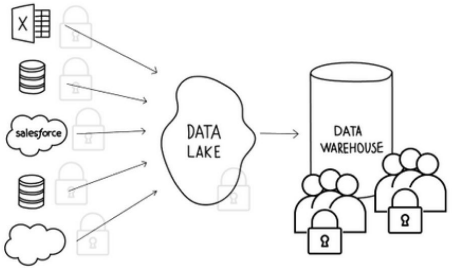
Identity and Access Management: permisos, roles, identidades

Cuentas de servicio: sin password, public/private key

Encripción: at rest, in transit

Key management: administrador por el cloud o por uno mismo

Legal compliance: HIPPA, COPPA, FEDRAMP, GDPR

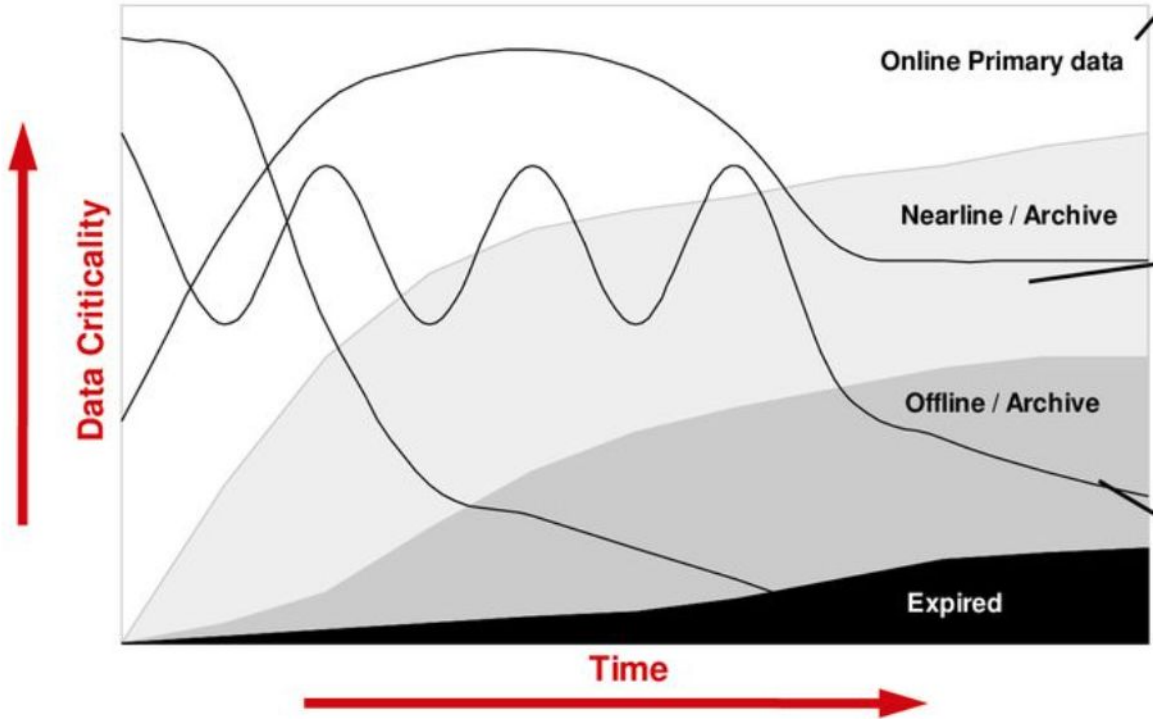


fuente: dataschool.com

Almacenamiento

Tiempo de almacenamiento

- Online data
- Nearline data (30 días)
- Offline (360 días)
- Expired (borrada)



Ciclo de vida

Proceso & Análisis

Proceso y Análisis

Data transformations

- Limpieza
- Lógica de negocio
- Filtrado
- Estandarizar data

Proceso y Análisis

Data analysis

Extraer información valiosa mediante técnicas:

estadísticas -> campos numéricos

texto -> campos no numéricos

Proceso y Análisis

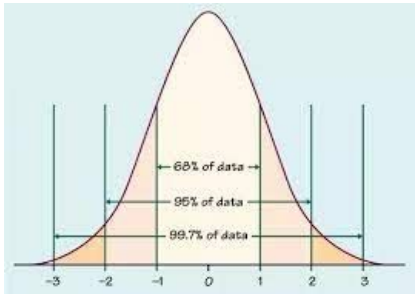
Data analysis

Extraer información valiosa mediante técnicas:

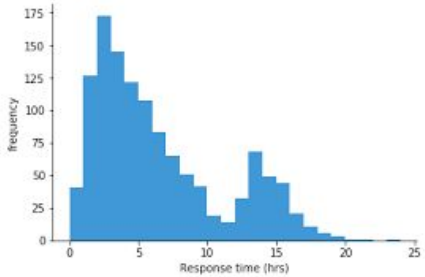
estadísticas -> campos numéricos

texto -> campos no numéricos

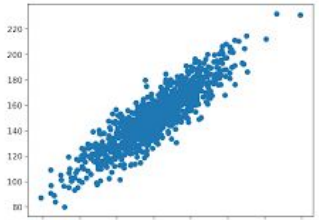
Media / desviación estándar



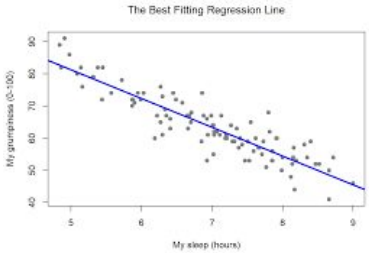
Histogramas



Correlación entre variables



Modelos de regresión



Proceso y Análisis

Data analysis

Proceso y Análisis

Data analysis

Extraer información valiosa mediante técnicas:

estadísticas -> campos numéricos

texto -> campos no numéricos

Contar la ocurrencia de una palabra determinada

Extraer entidades (direcciones, nombres, num. teléfonos, etc.)

Ciclo de vida

Visualización & exploración

Exploración

- Explorar información y modelos de ML
- Ambiente listo para usar
- Compartir facilmente
- Lenguajes: Python, R, SCALA
- Gratuito
- Uso de Hardware disponible en la nube (GPU, TPU)

Exploración

red neuronal en Python y Tensorflow

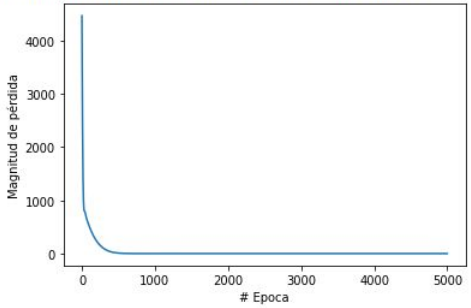
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Última modificación: 17 de febrero

Comentario Compartir

+ Código + Texto Conectar Editar

```
[ ] import matplotlib.pyplot as plt
    plt.xlabel("# Epoca")
    plt.ylabel("Magnitud de pérdida")
    plt.plot(historial.history["loss"])
```

```
[<matplotlib.lines.Line2D at 0x7ff527bdbdd0>]
```




```
[ ] print("Hagamos una predicción!")
    resultado = modelo.predict([-40])
    print("El resultado es " + str(resultado) + " fahrenheit!")
```

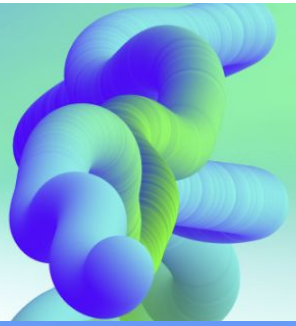
```
Hagamos una predicción!
El resultado es [[-39.96545]] fahrenheit!
```

Exploración

Curso de Data Science con Python

Este curso de data science con Python te permitirá adentrarte en el mundo de los patrones, desde una mirada intuitiva, con guías paso a paso tanto en la teoría como en la práctica utilizando el lenguaje Python.

Es un curso práctico, orientado a que puedas desarrollarte como **Data Scientist Jr**.

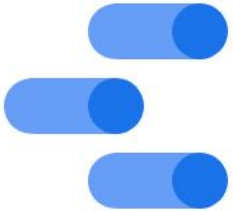


Introducción a Python

Aprende lo esencial en 15 días con nuestro curso de introducción a Python. Domina los conceptos básicos para comenzar a trabajar con Datos en Python.

Tendrás a **Atenea nuestra IA** que te guiará en tu aprendizaje en todo momento y también contarás con soporte de nuestro instructores a través de nuestra comunidad.

Visualización



Visualización

- Permiten crear distintos tipos de visualización (Dashboards, KPIs, reportes, análisis)
- Permite crear distintos tipos de gráficos (treemap, torta, línea, barra, maps, etc.)
- Distintas fuentes de datos (CSV, TXT, Excel/Sheet, DBs, etc.)
- Leer información online - offline
- Permite fácilmente compartir entre equipos de trabajo
- Gratuitos y pagos

Aspectos técnicos y estructuras

Aspectos técnicos de la data

- Volumen: Terabytes hasta Petabytes de información
- Velocidad: cuán rápido está ingresando la información para almacenarla
- Variación: cantidad de variación en la estructura de la información
- Veracidad: información precisa, limpia y veraz
- Valor: capacidad de convertir mucha información en valor para el negocio
- Acceso: como la información va a ser leída o escrita y cada cuanto tiempo.
- Seguridad: contemplar los distintos tipos de acceso a la información

Tipos de estructura

- Estructurada
- Semi estructurada
- No estructurada