Clase 4 IA - BigData

Tipos de datos

- Datos Estructurados: A menudo numeroso etquetas almacenadas en un marco estrucuturados de columnas y filas.
- Datos Semiestructurados: Organizados libremente en categorias utiilizando etiquetas meta.
- Datos No Estructurados: Infomracion co muicho texto que no esta organizada en u marco o modelo claramente definido.

ETL vs ELT

ETL

- Extract: Se obtienen los datos desde diversas fuentes (api, db).
- Transform: Los dats extraidos son procesados y transformados fuera del sistema destino. Se limpian, formatean y estrcturan de acuerdo a las reglas de negocio.
- Load: Los datos transformados se cargan en un almacen.

ELT

- Extract: Los datos son extraidos de las fuentes, igual que en ETL
- Load: Los datos sin transformar se cargan directamente en el sistema destino. > Es un concepto alimentado por la nube
- Transform: La transformacion ocurre dentro delo sistema de destino, aprovechando su capacidad de procesamiento.

Categorias

Bronce

Tener los datos en bruto, por ejemplo, cuando hacemos web-scraping. a veces es json, csv. masomenos estructurados.

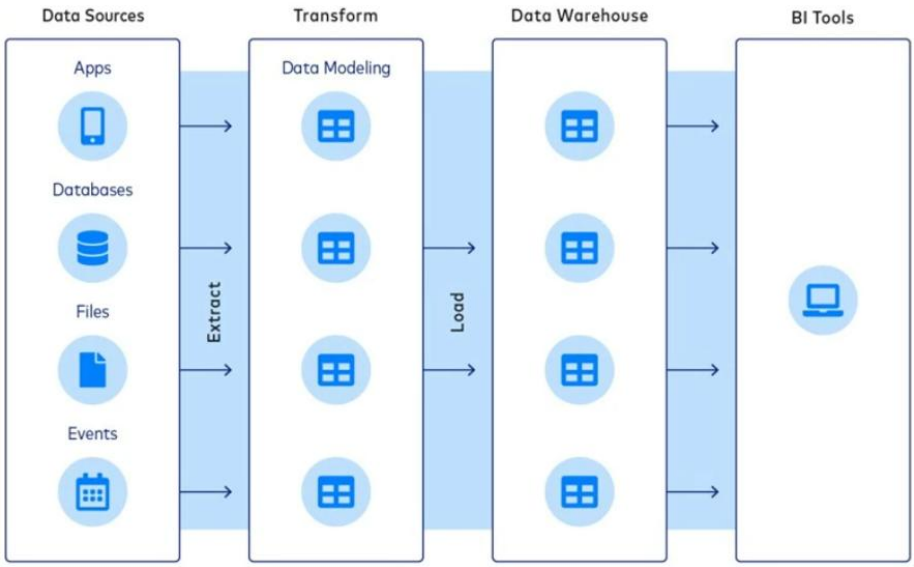

Figure 1: Imagen de referencia

Plata

es el procesamiento de los datos de tipo bronce. simplemnte una pequeña modificacion de los datos que los deje listos para hacer una transformacion mas completa.

Oro

Es el dato más refinado. Esto es luego de hacer agregaciones (porque necesitan mucho computo). Este es el nivel desde donde deberia consumir el sistema

Herramientas que se usan: > Azure, Apache spark.

Webscrapping

Es un procedo mediante el cual extraemos datos de la web de forma medianamente automatizada.

Nota

escribir la direfencia entre **Data Factory** contra **Data Bricks**.