Apuntes #1

Bases de Datos II

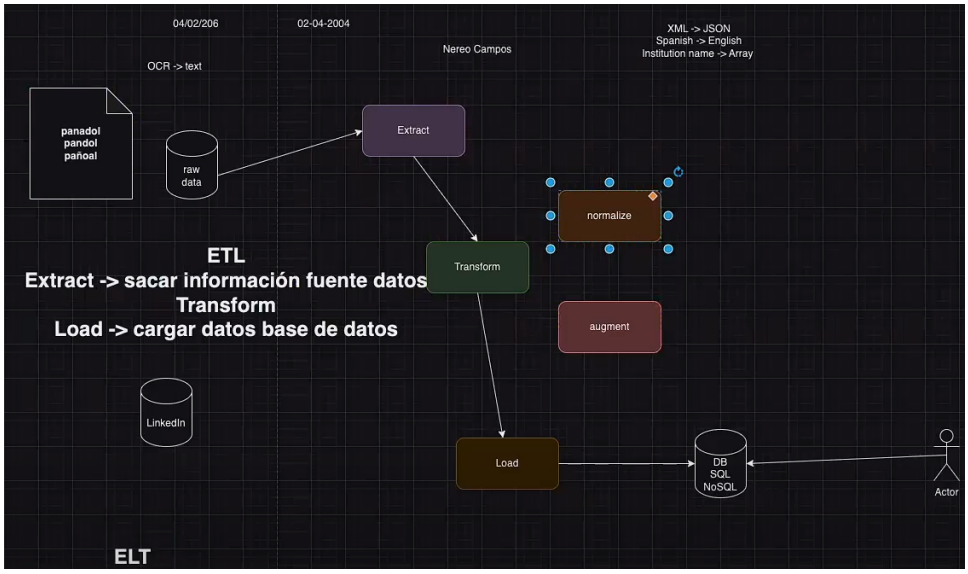
Luis Diego Delgado Muñoz

Prof. Nereo Campos

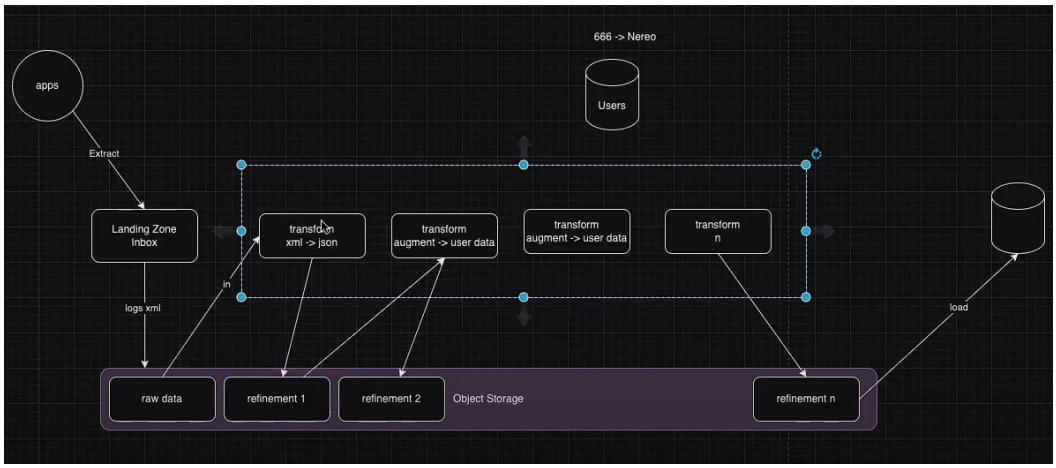
Clase del 16/02/2024

- SparkQL se utilizará para el proyecto opcional.
- **Raw data o datos crudos:** Datos que se recolectan específicamente de un sistema, son datos que no han sido procesados.

  Ejemplo: si se agarra un libro y lo digitalizamos, lo hacemos pasar por un proceso de OCR, generará texto sobre las páginas que son imágenes se podrán guardar como datos crudos que no han sido procesados.
- Pasos para realizar un procesamiento de datos:
  - Normalizar datos
  - **Aumentar los datos:** derivar datos a partir de datos ya existentes.
  - **Transformación de datos:** cualquier cambio que se le realice a los datos, ya sea cambiar de formato o traducción de datos.
- **Load:** consiste en cargar datos desde una fuente de datos cruda y agregarlos en el sistema pipeline de transformación de datos y será enviada a una base de datos.
- **Extract-transform-load (ETL)**
  - No siempre es cargado a una base de datos, pues en algunas ocasiones las empresas usan datos curados.
  - No es necesario hacer el ETL en ese orden, se puede usar un ELT, lo cual consiste en extraer, cargar en un sistema de base de datos y se transforman. **Ejemplo:** extraer datos, cargarlos a elasticsearch o en una base de datos relacional, como Data warehouse, después el sistema te permite transformar el dato una vez cargado.

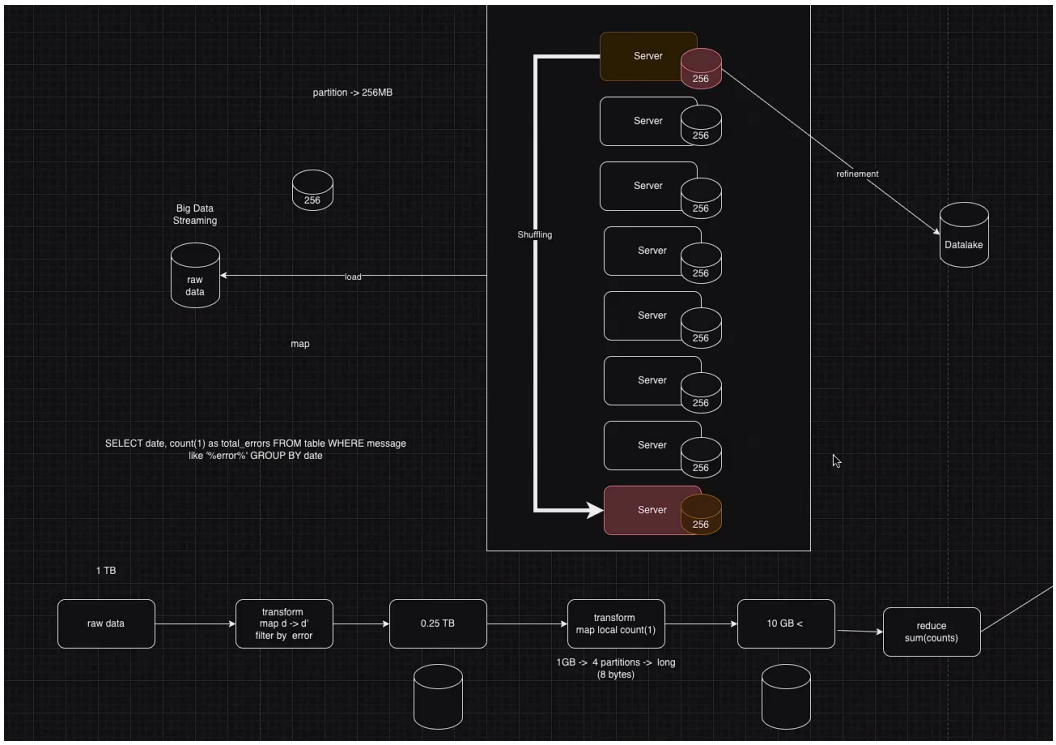


- **Data lake:** ubicación en la nube con almacenamiento masivo.
- Los cloud providers tienen conceptos llamados **Object storage**, el cual es un lugar infinito (masivo) de almacenamiento.
- Los **Objects storage** nos permite tener landing zone o Inbox el cual es un sistema, app o persona que realiza una publicación de datos (Logs) los cuales son RAW data. Se implementa un pipeline de transformaciones para cambiar formatos, lo cual leerá el inbox y tendrá una sección que se llama refinements donde se guardan los datos de esta transformación y de igual forma es la entrada a otras transformaciones, y así de manera que el data lake guarde información masiva, donde después podremos publicar en un sistema externo.

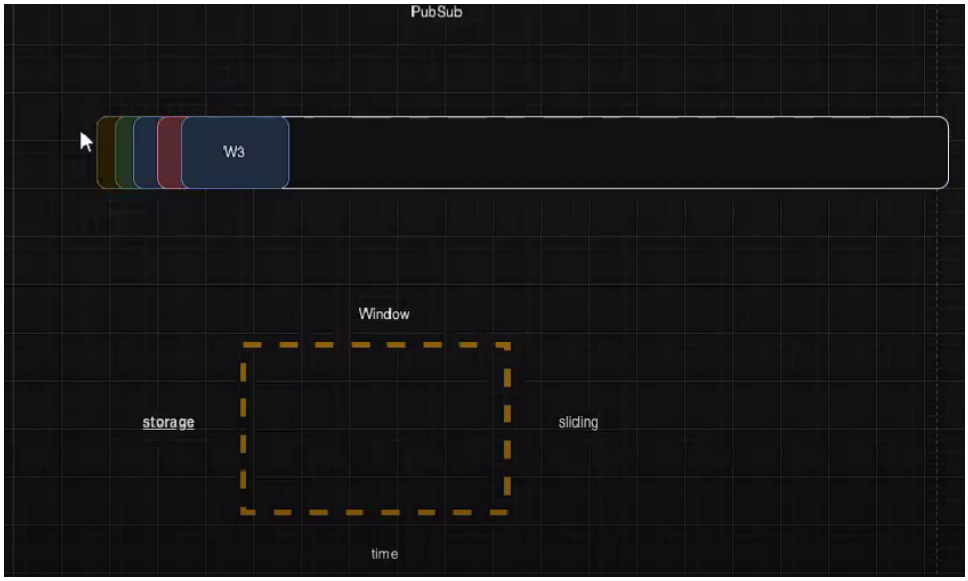


- **Spark** Es un sistema distribuido, es decir un grupo de máquinas o computadoras que van a coordinar de alguna forma. Normalmente en este tipo de sistemas las computadoras se van a orquestar de alguna manera trabajando en paralelo.

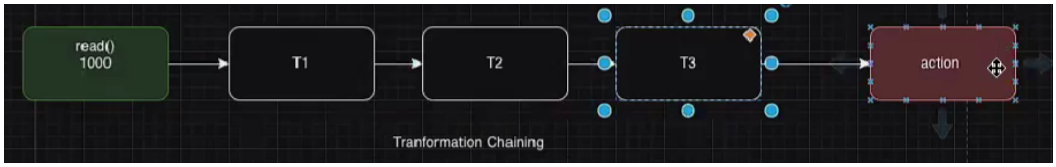
- **Framework** Raw data envía acciones al framework para que realice, el framework será la carga (load) de los datos, estos datos serán tan masivos que la carga de datos no entre en un servidor. Framework creara particiones, las cuales se encargan de dividir los megas en 256MB entre cada servidor para realizar el procesamiento. Existen procesamientos que se pueden realizar en paralelo y otros que no
- **Shuffling:** intercambio de datos entre los servidores, resulta como problemática porque no es eficiente para los servidores.



- **Internet of things (IoT)**
  - Source data genera un flujo de información infinito.
  - Los datos deben entrar a un sistema de almacenamiento de datos temporal
- Sistema estándar de la industria **Apache Kafka**, el cual es un hibrido entre una base de datos y un sistema de mensajería, permite almacenar el tiempo necesario los datos.
- **PubSub**
  - Window, puede tener diferentes configuraciones
    - Time
    - Storage
    - Sliding



- **Transformation chaining**



- **Action:** es lo que desencadena el procesamiento, en el momento en el que se leen los datos la cadena de transformación hasta el momento momento en el que se aplica el action.

Nota: se revisó el código de Scala en clase.