

# Curso Data Engineer: Creando un pipeline de datos

MÓDULO C - Clase 5

A decorative graphic on the left side of the slide, consisting of a grid of small dots in various shades of blue and green, arranged in a pattern that tapers towards the top and bottom.

# GCP Ingest

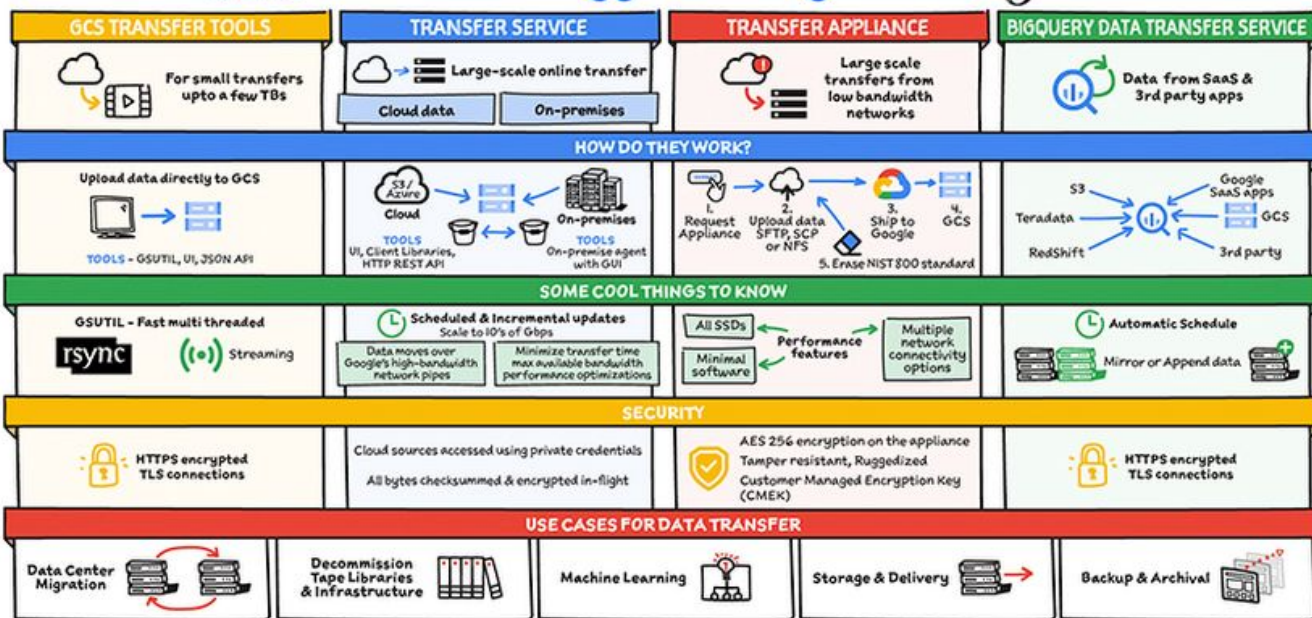
# Ingest



#GCPSSketchnote  
@PYERGADIA  
THECLOUDGIRL.DEV  
03.30.2021



## Options to move data to Google Cloud



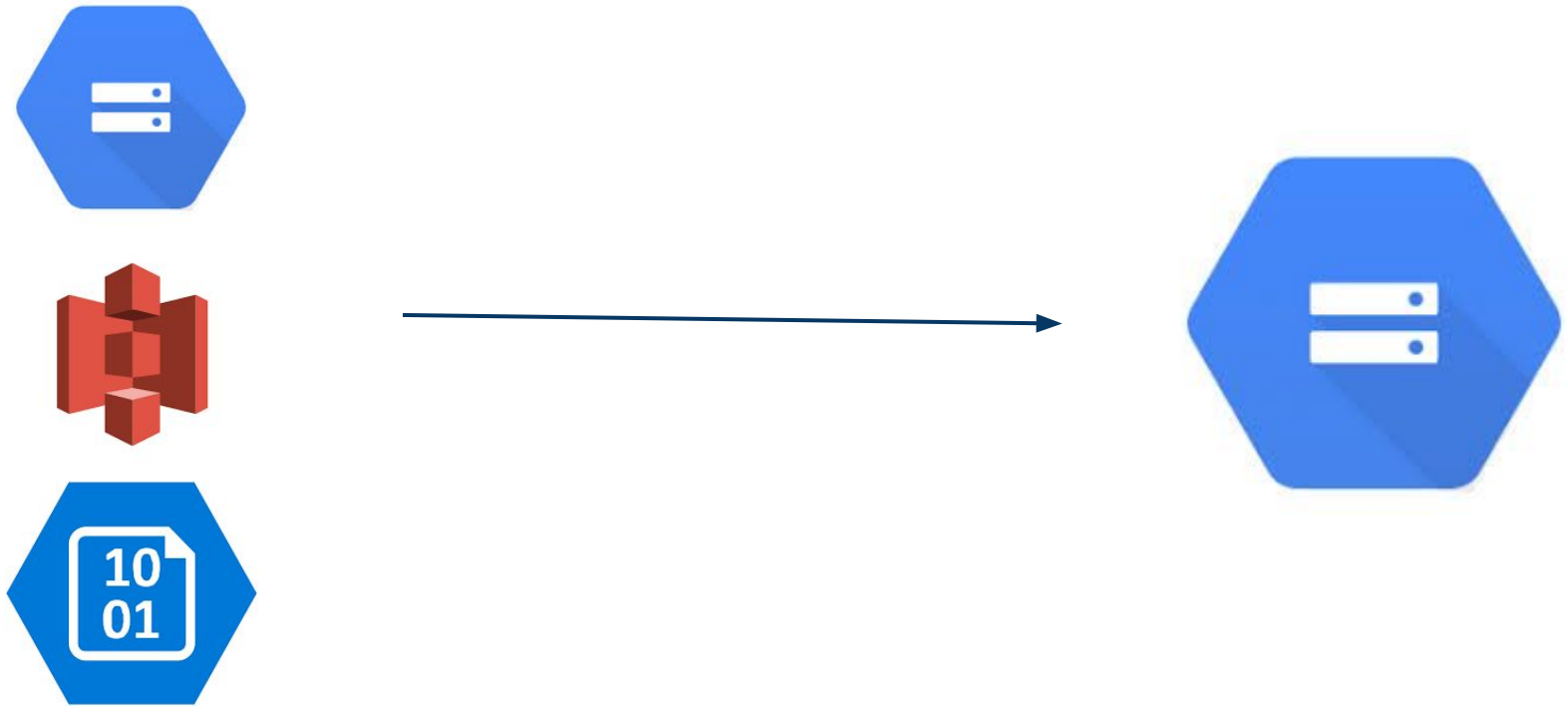
gsutil



 Parquet



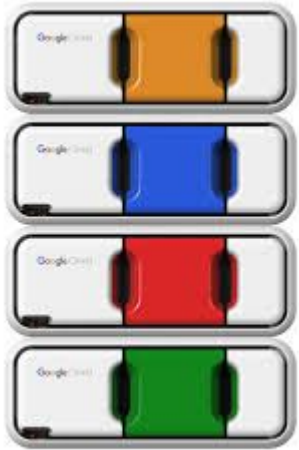
# transfer service - data transfer



# transfer appliance



# transfer appliance





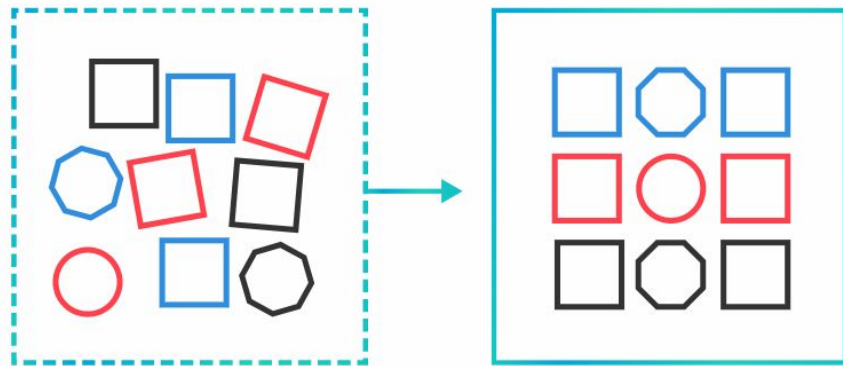
Transform



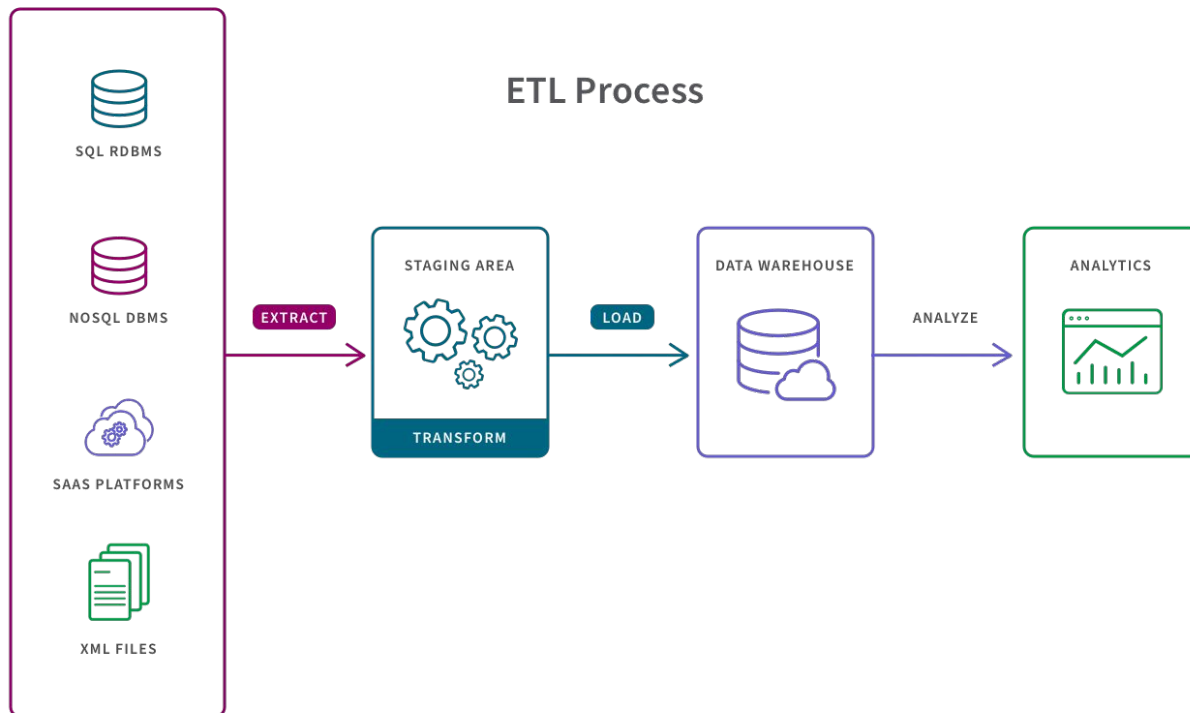
# Transform



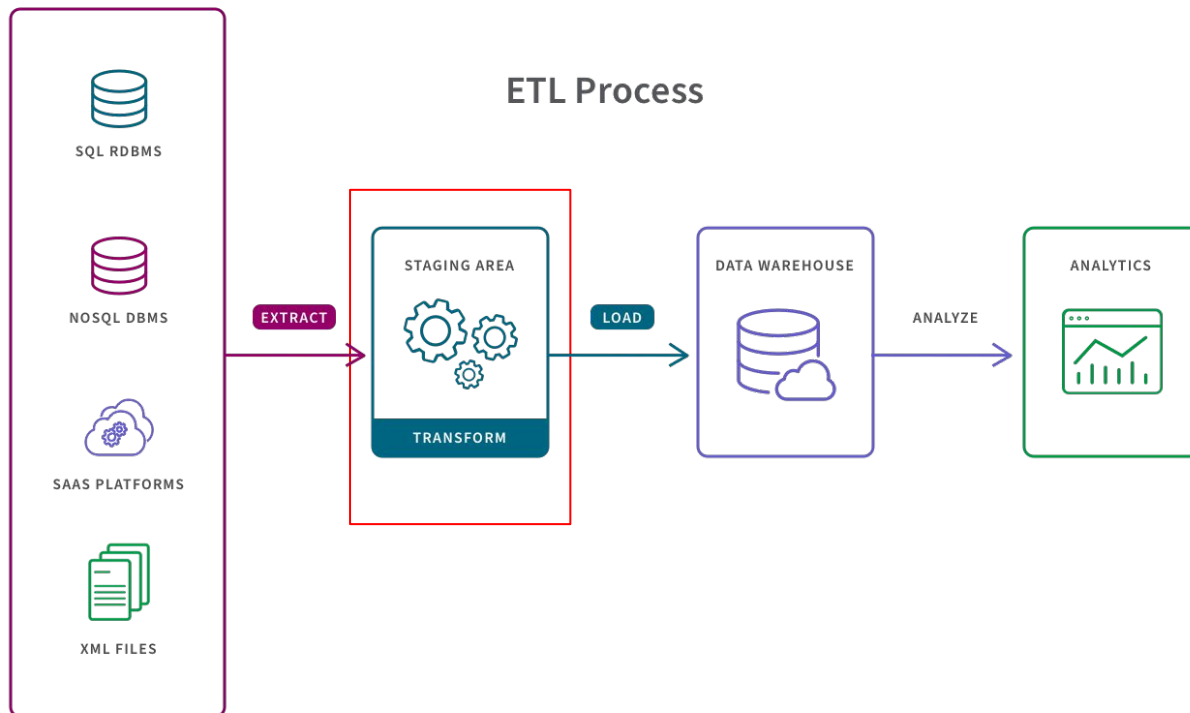
Es el proceso de convertir, limpiar y estructurar datos en un formato utilizable que se pueda analizar para respaldar los procesos de toma de decisiones.



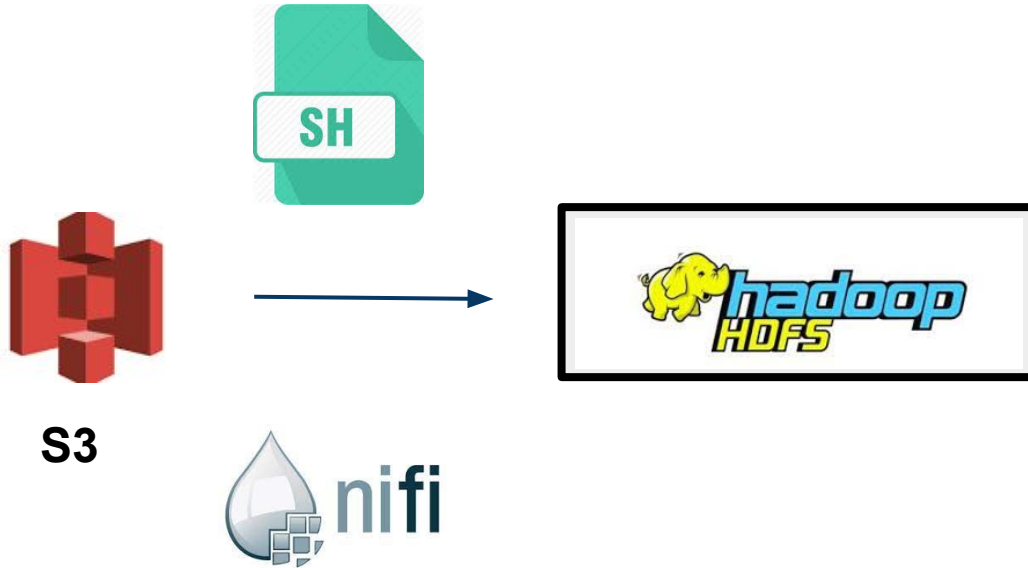
# Transform



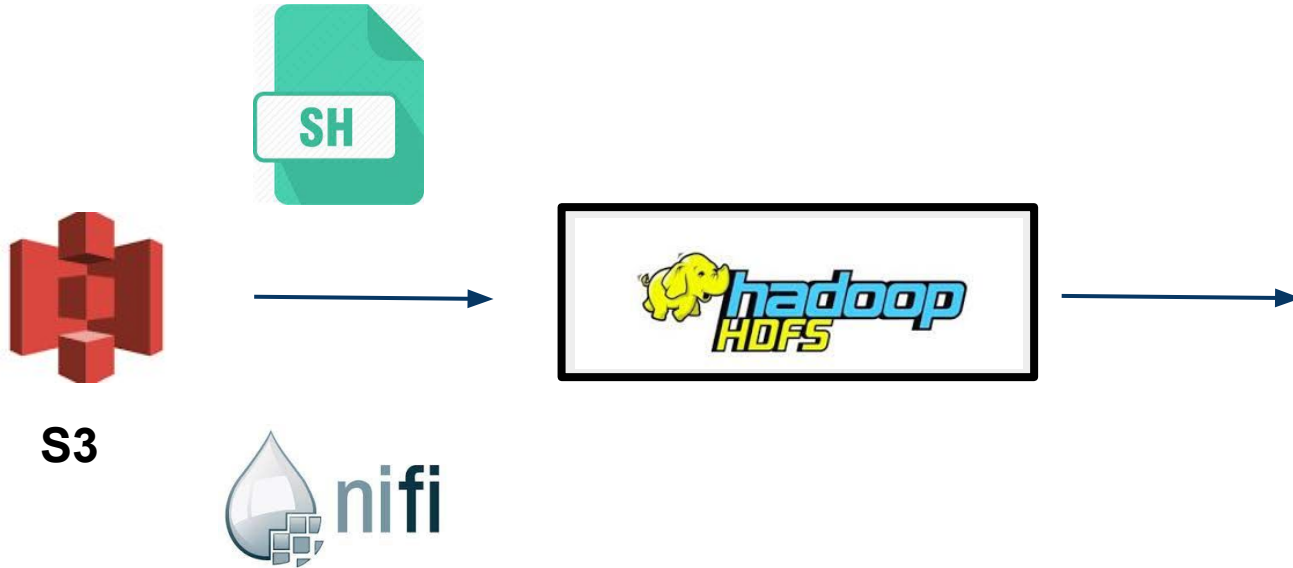
# Transform



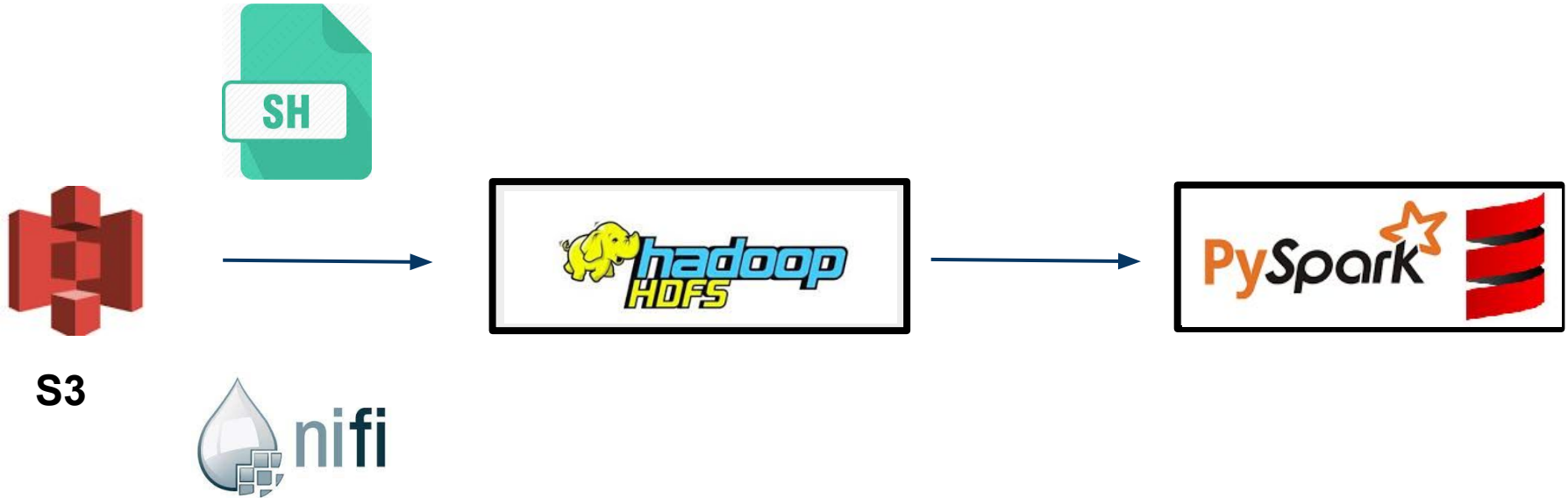
# Transform



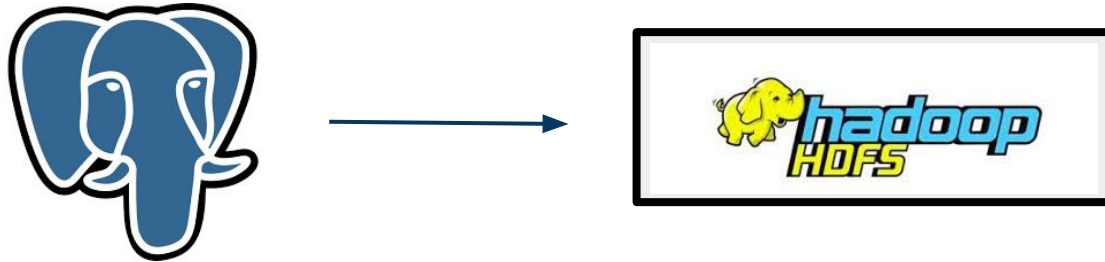
# Transform



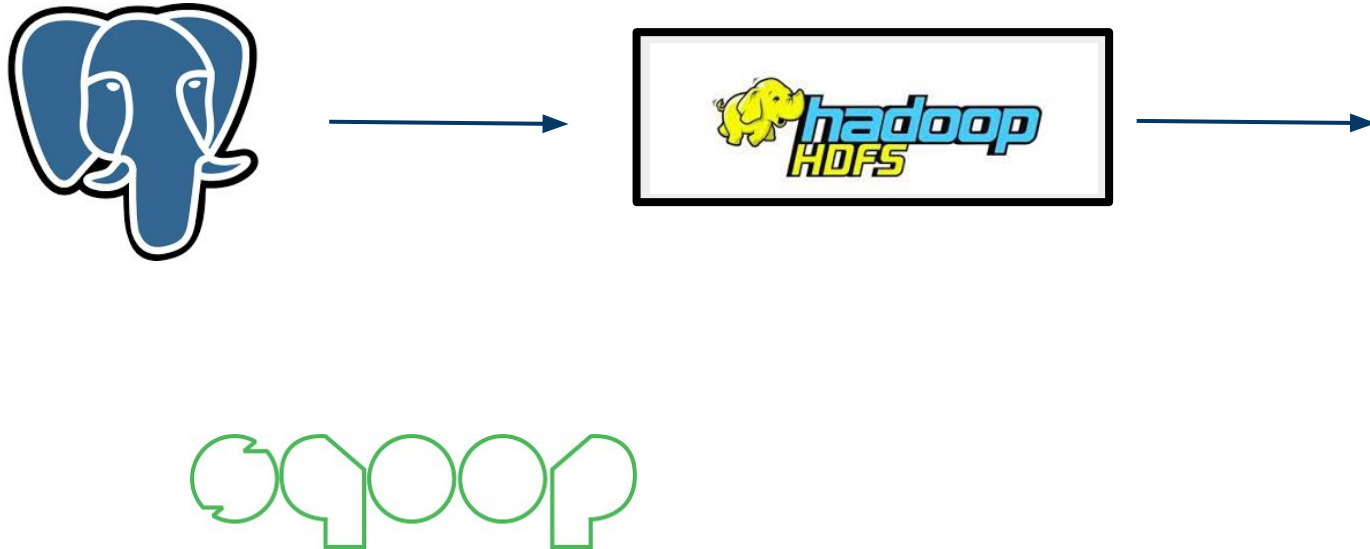
# Transform



# Transform



# Transform





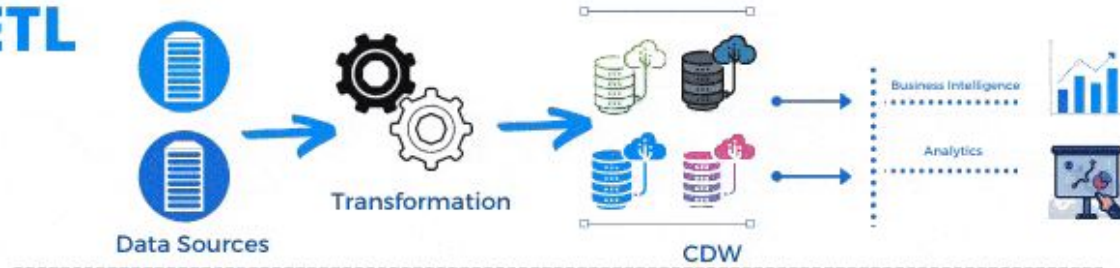
# Transform



# ETL VS ELT



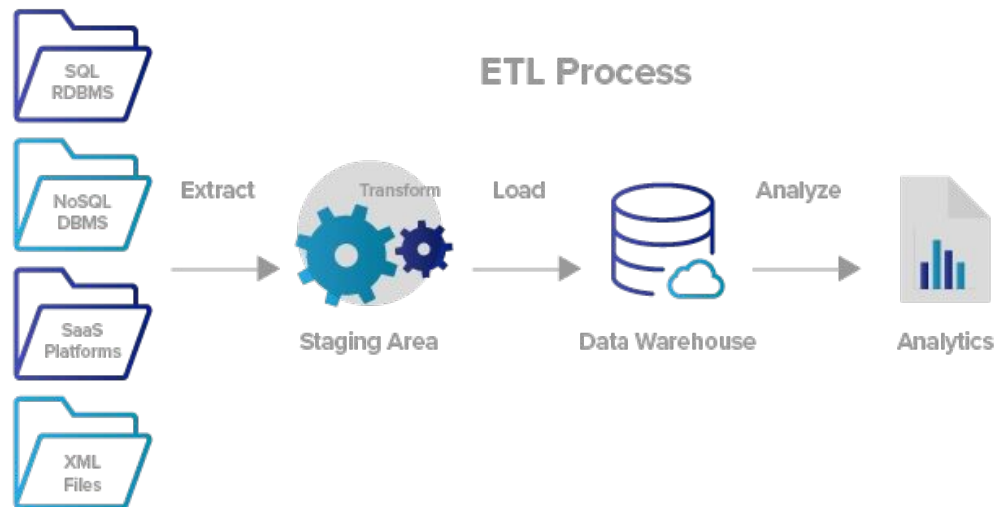
**ETL**



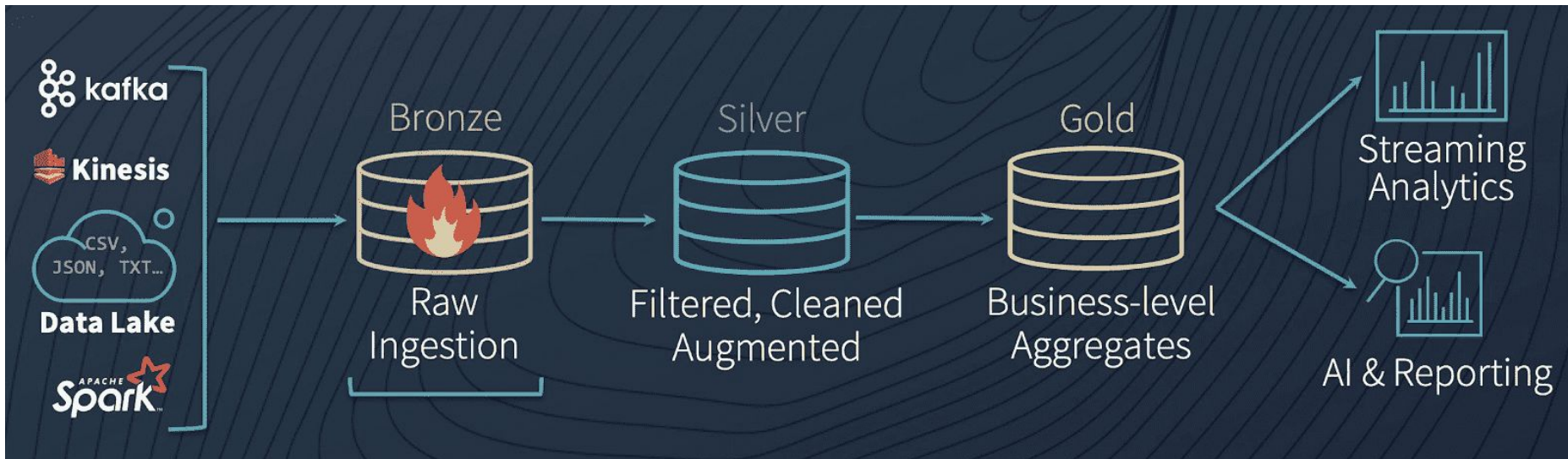
**ELT**



# ETL



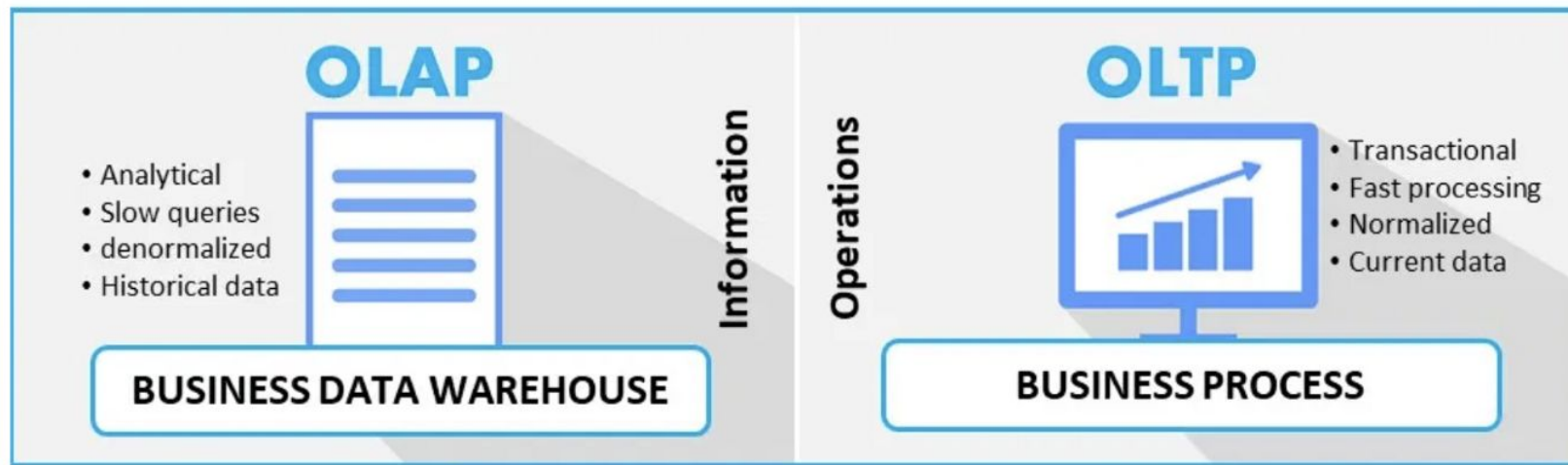
# ELT



## Online Analytical Processing Vs Online Transaction Processing



# OLAP Vs OLTP



# On line Analytical Processing Vs On line Transaction Processing



Students	
id	student_name
1	Juan García
2	José Perez
3	Alberto Quiroga

Courses	
id	course_name
1	SQL
2	Python
3	R
4	Java

Students_Courses			
Date	id_student	id_courses	mark
11/01/2022	1	1	7
10/25/2022	1	2	6
10/28/2022	2	4	6
10/03/2022	3	3	5

# ETL



1. Con Sqoop hago un export en archivos parquet/text/avro, etc.
2. Ingesto esos archivos en HDFS (hdfs dfs -put origen destino)
3. Transformo esos archivos con pyspark/scala/sql
4. Cargo esos datos en el Data Warehouse (Hive)

# OLAP



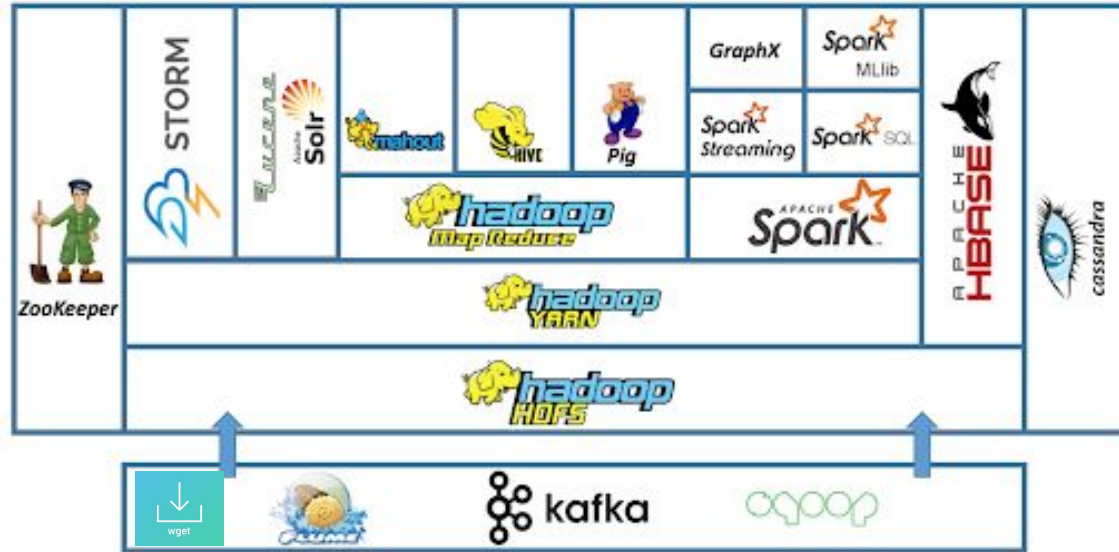
Marks				
Date	student	courses	mark	assist
11/01/2022	Juan García	SQL	7	100
10/25/2022	Juan García	Python	6	93
10/28/2022	José Perez	Java	6	95
10/03/2022	Alberto Quiroga	R	5	80





# Ejercicio

# Ecosistema Hadoop



# Ejercicios



- Transform
  - Transformaciones en Pyspark (SQL)
  - Transformaciones en Scala