Escuela de Datos Vivos

Curso Data Engineer: Creando un pipeline de datos

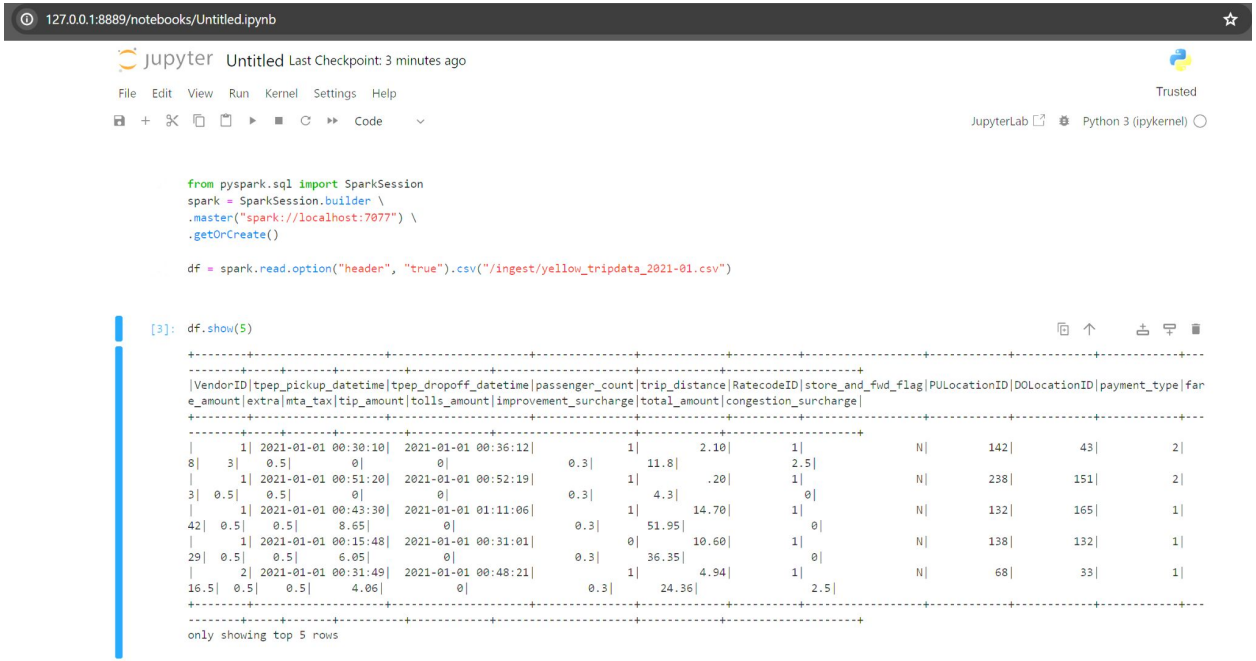
MÓDULO B - Clase 5

Agenda

- Jupyter Notebook
- GCP Ingest
- Transform
- Ejercicios

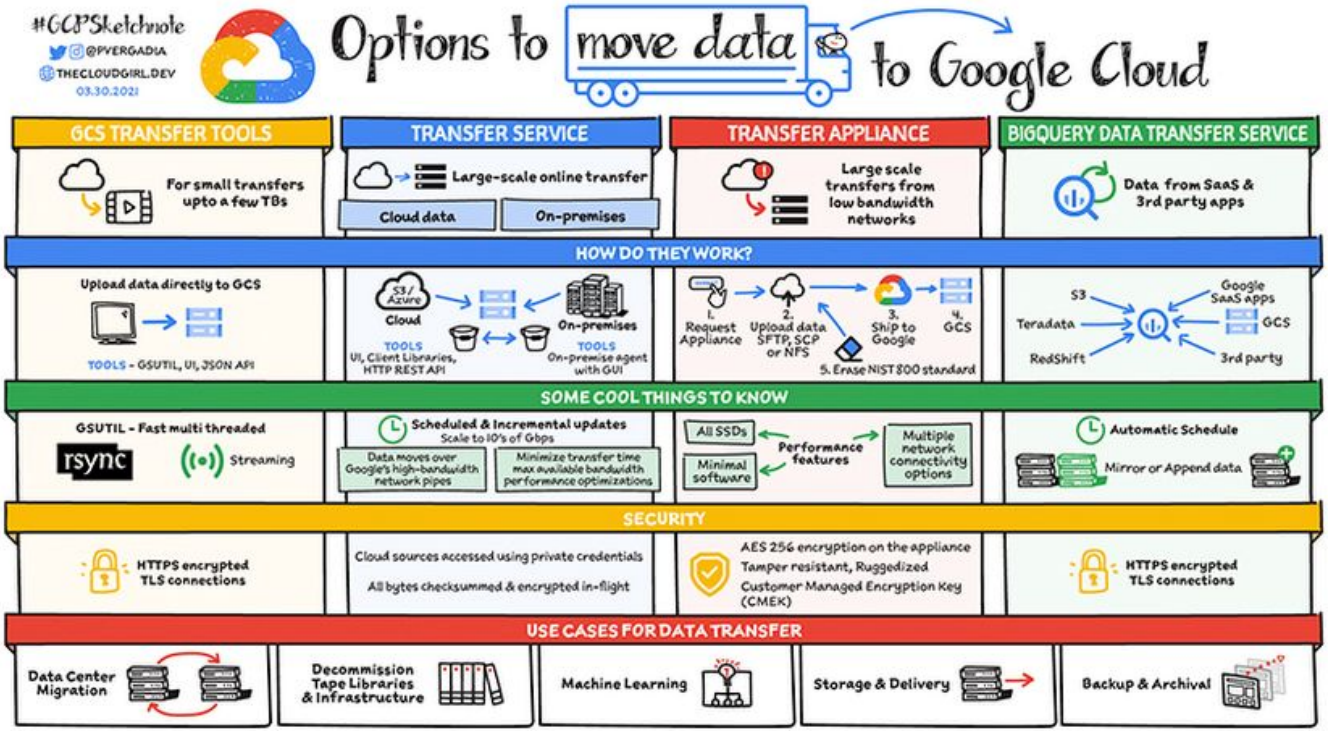
Jupyter Notebooks

Jupyter Notebook



GCP Ingest

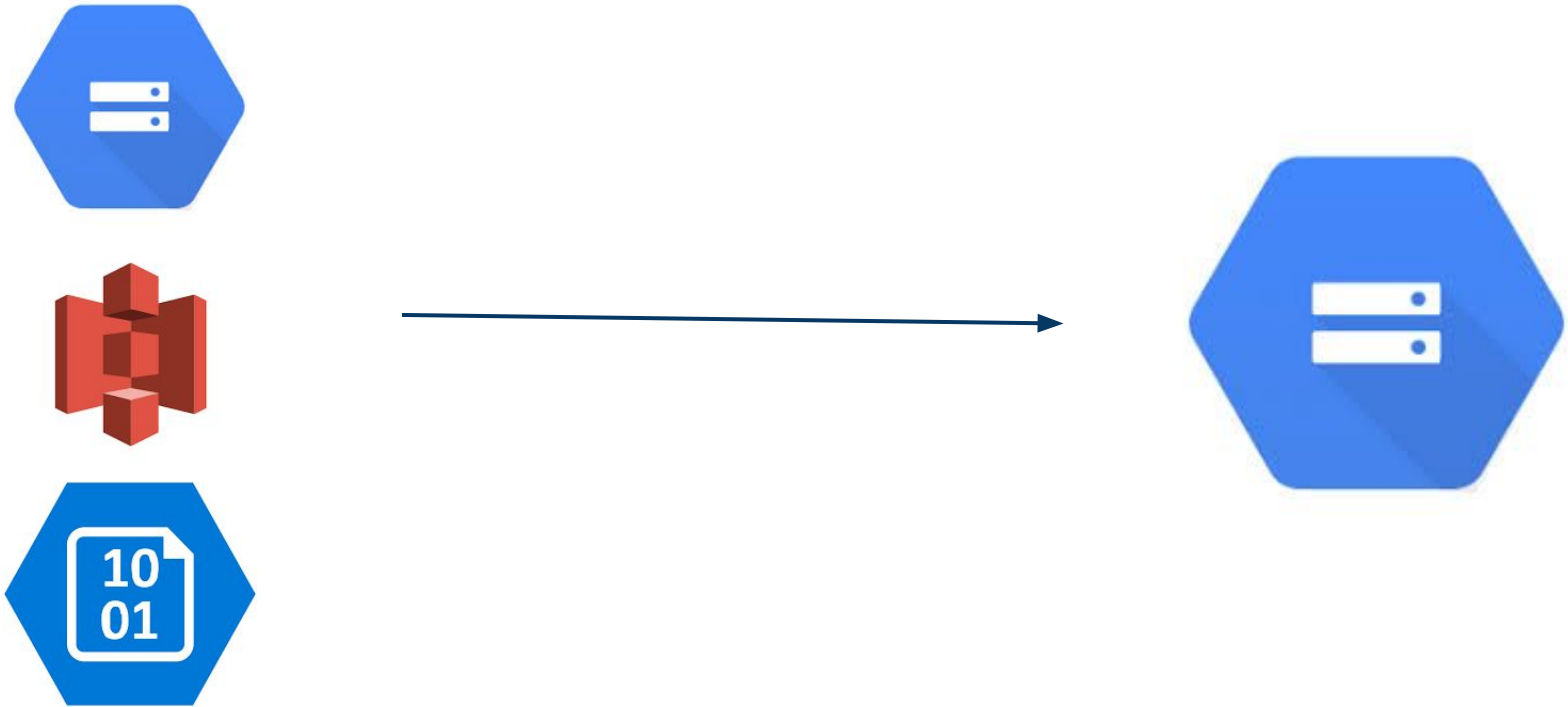
Ingest


#GCPSketchnote
@PVERGADIA
THECLOUDGIRL.DEV
03.30.2021
Options to move data to Google Cloud
GCS TRANSFER TOOLS
For small transfers upto a few TBs
TRANSFER SERVICE
Large-scale online transfer
Cloud data
On-premises
TRANSFER APPLIANCE
Large scale transfers from low bandwidth networks
BIGQUERY DATA TRANSFER SERVICE
Data from SaaS & 3rd party apps
HOW DO THEY WORK?
Upload data directly to GCS
TOOLS - GSUTIL, UI, JSON API
S3 / Azure
Cloud
On-premises
TOOLS
UI, Client Libraries, HTTP REST API
TOOLS
On-premise agent with GUI
1. Request Appliance
2. Upload data SFTP, SCP or NFS
3. Ship to Google
4. GCS
5. Erase NIST 800 standard
S3
Teradata
RedShift
Google SaaS apps
GCS
3rd party
SOME COOL THINGS TO KNOW
GSUTIL - Fast multi threaded
rsync
Streaming
Scheduled & Incremental updates
Scale to 10's of Gbps
Data moves over Google's high-bandwidth network pipes
Minimize transfer time max available bandwidth performance optimizations
All SSDs
Minimal software
Performance features
Multiple network connectivity options
Automatic Schedule
Mirror or Append data
SECURITY
HTTPS encrypted TLS connections
Cloud sources accessed using private credentials
All bytes checksummed & encrypted in-flight
AES 256 encryption on the appliance
Tamper resistant, Ruggedized
Customer Managed Encryption Key (CMEK)
HTTPS encrypted TLS connections
USE CASES FOR DATA TRANSFER
Data Center Migration
Decommission Tape Libraries & Infrastructure
Machine Learning
Storage & Delivery
Backup & Archival

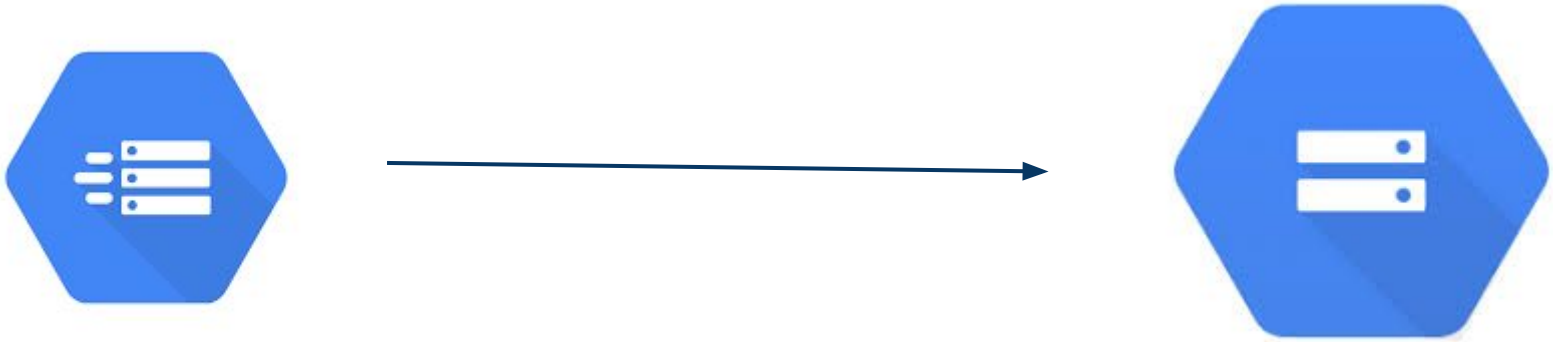
gsutil

Parquet

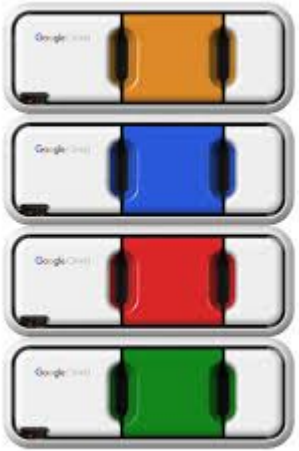
transfer service - data transfer



transfer appliance



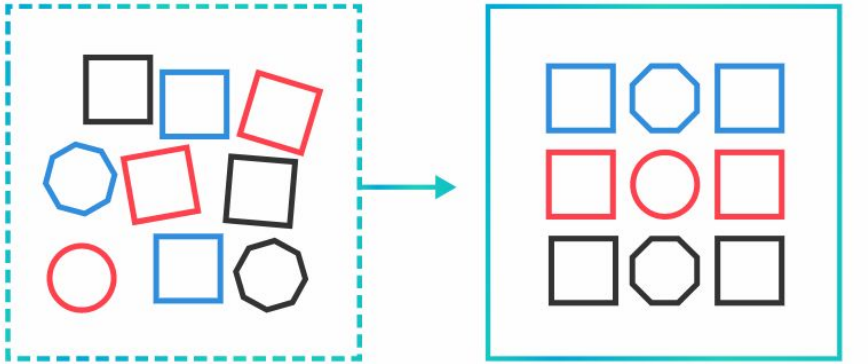
transfer appliance



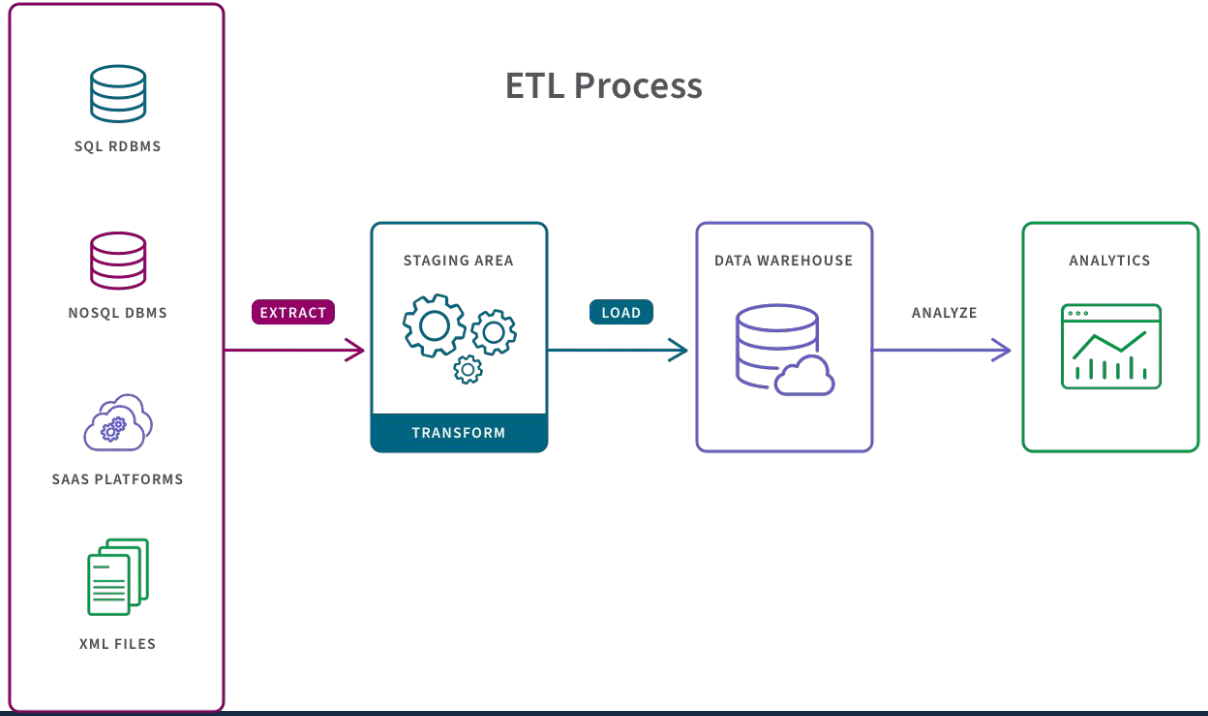
Transform

Transform

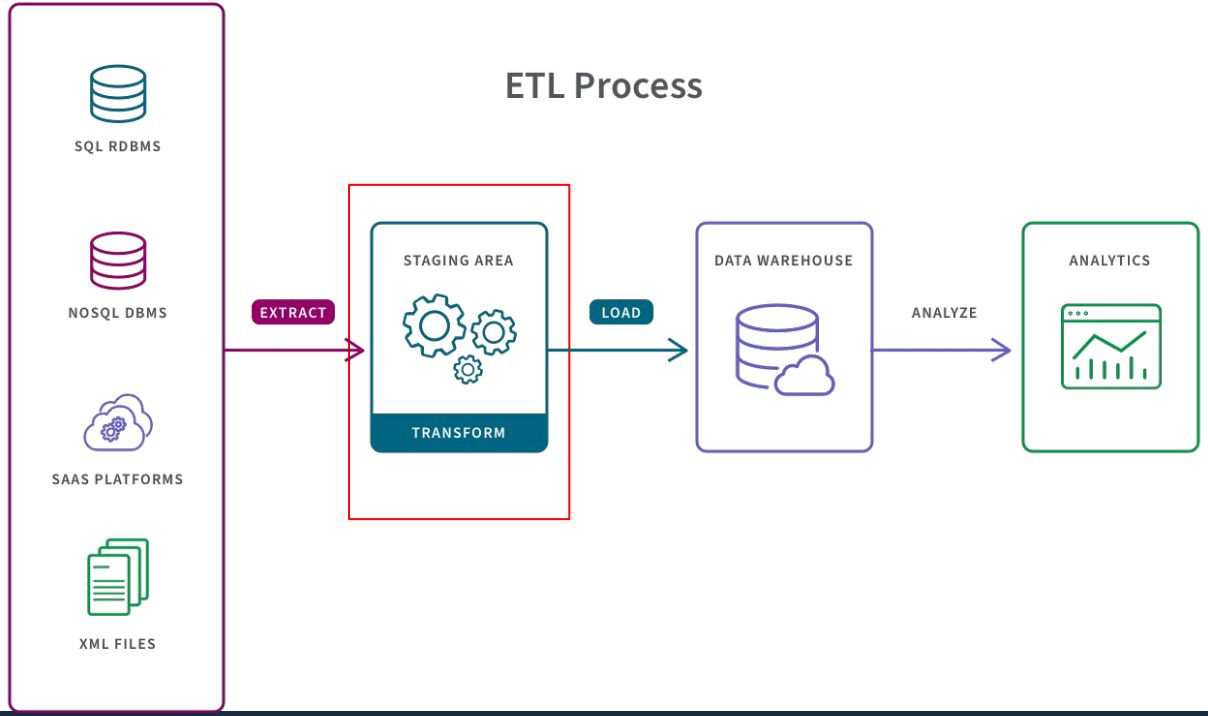
Es el proceso de convertir, limpiar y estructurar datos en un formato utilizable que se pueda analizar para respaldar los procesos de toma de decisiones.



Transform


ETL Process
SQL RDBMS
NOSQL DBMS
SAAS PLATFORMS
XML FILES
EXTRACT
STAGING AREA
TRANSFORM
LOAD
DATA WAREHOUSE
ANALYZE
ANALYTICS

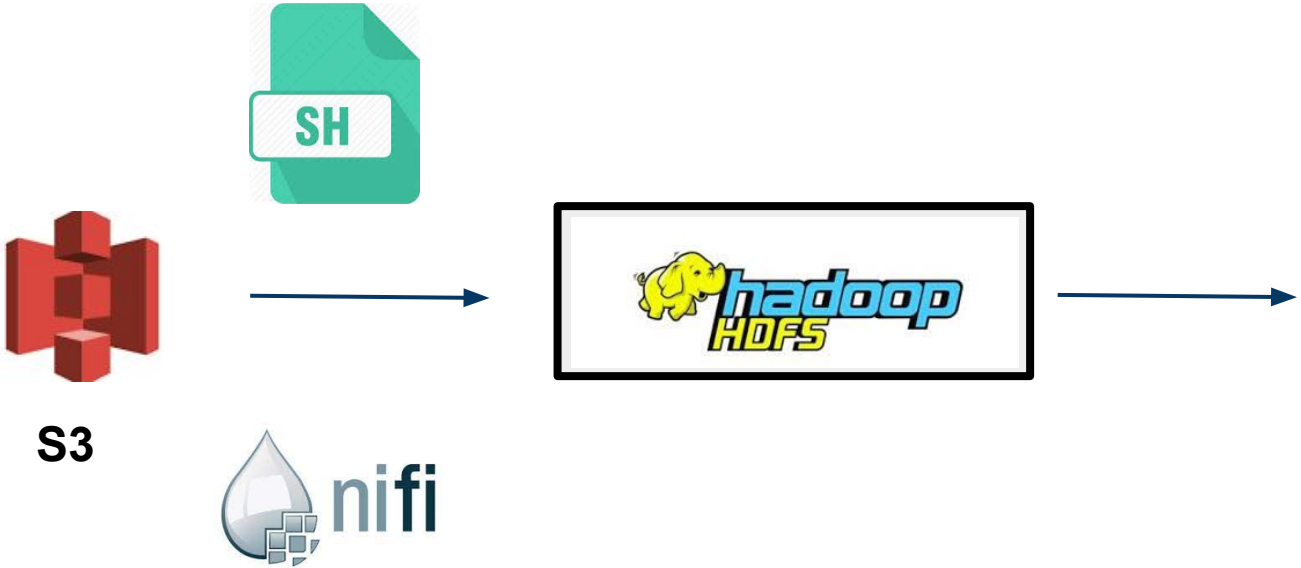
Transform


ETL Process
SQL RDBMS
NOSQL DBMS
SAAS PLATFORMS
XML FILES
EXTRACT
STAGING AREA
TRANSFORM
LOAD
DATA WAREHOUSE
ANALYZE
ANALYTICS

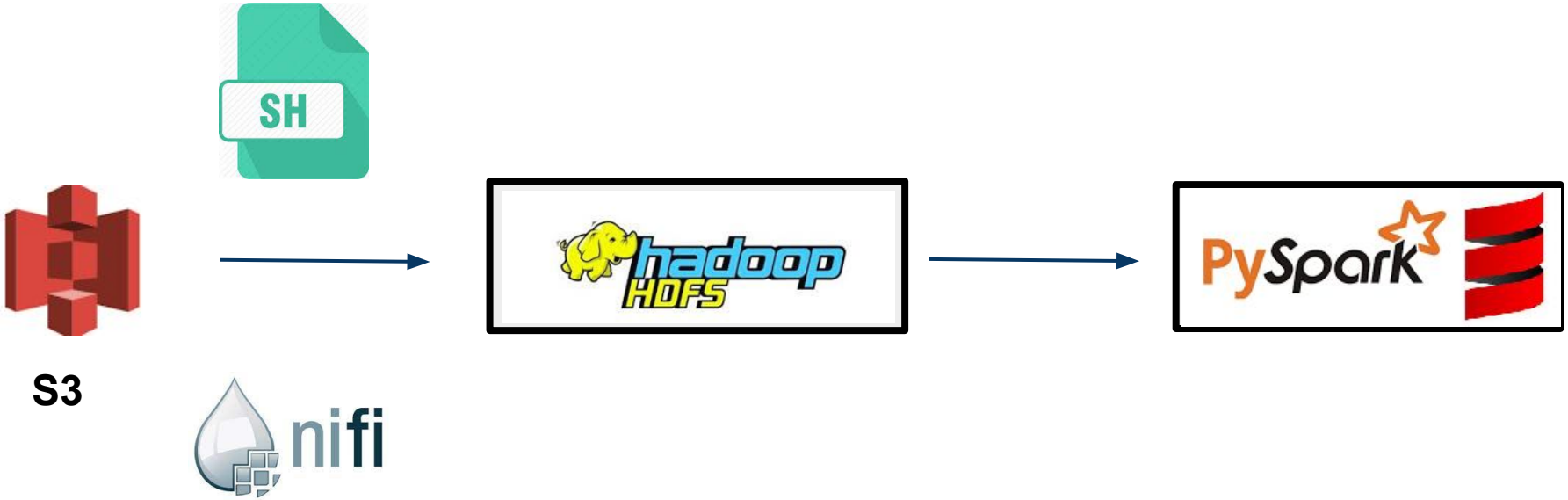
Transform

S3

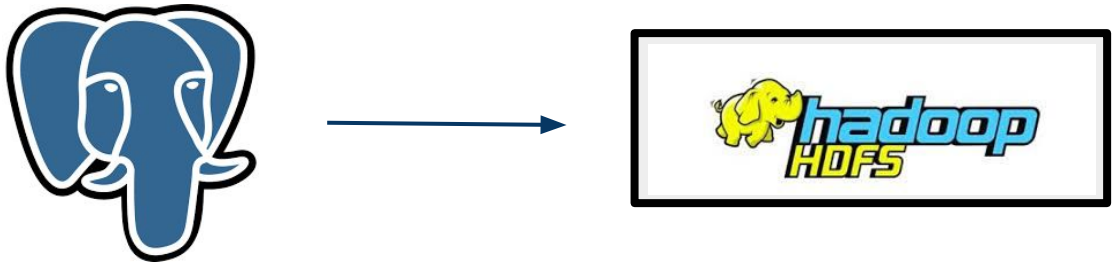
Transform



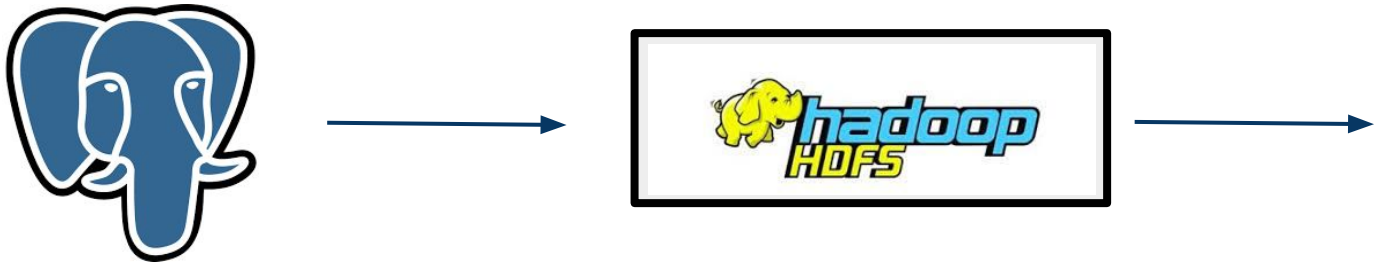
Transform



Transform

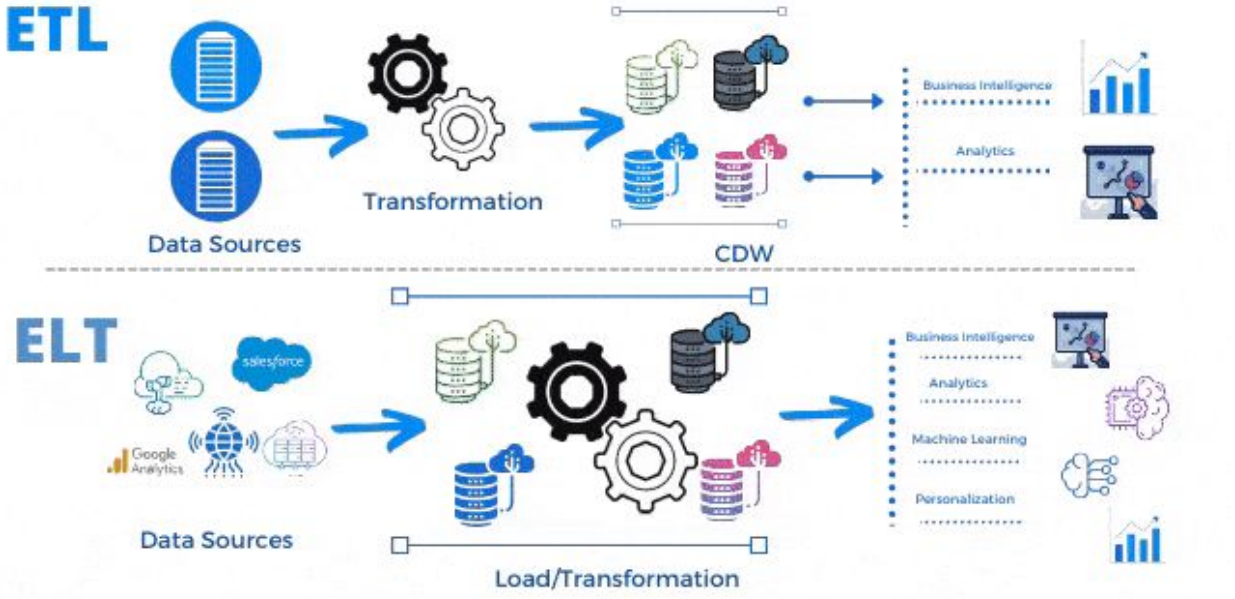


Transform

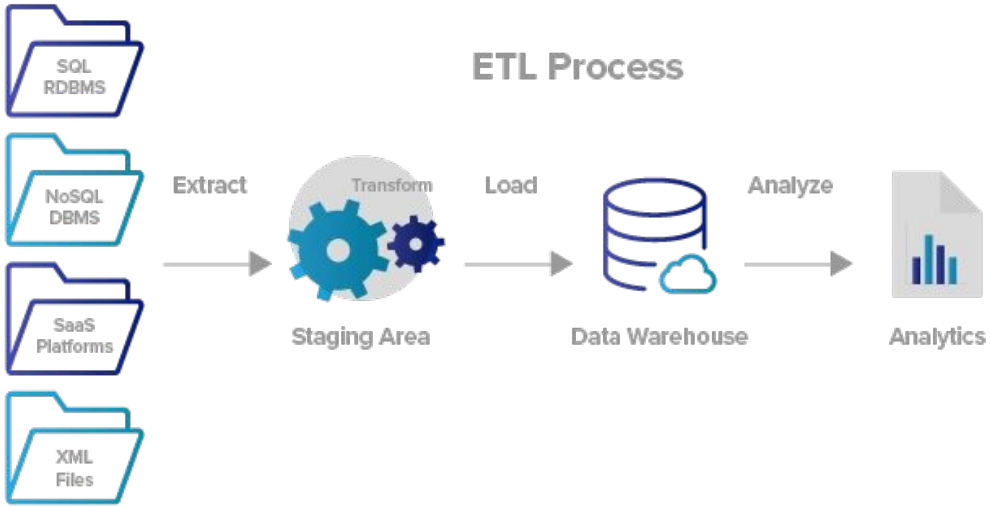


Transform

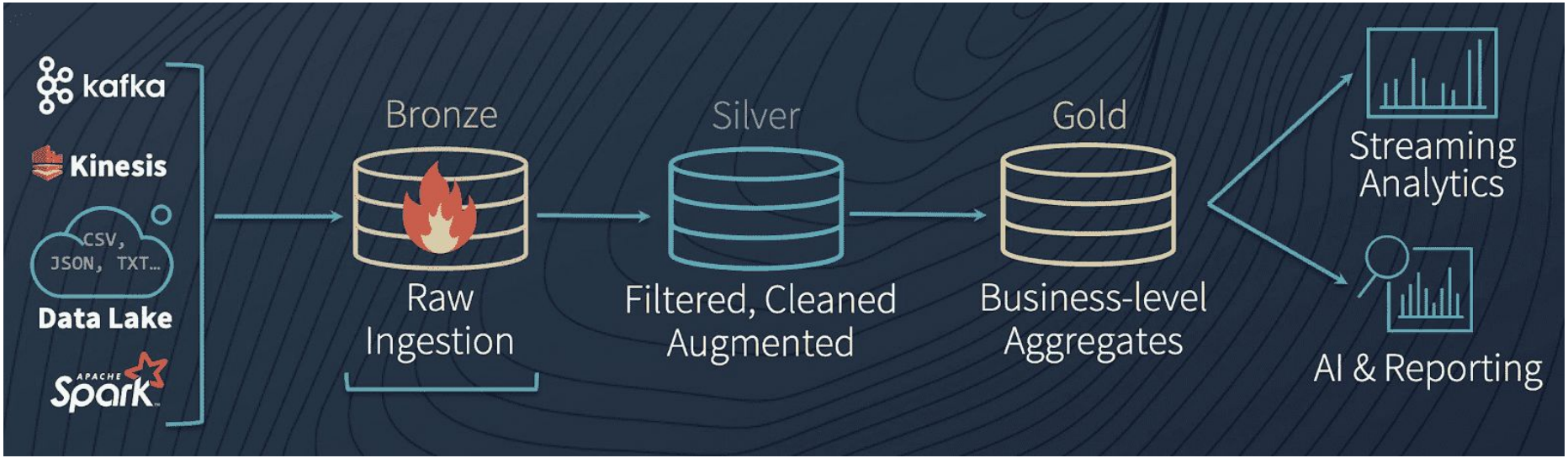
ETL VS ELT


ETL
Data Sources
Transformation
CDW
Business Intelligence
Analytics
ELT
salesforce
Google Analytics
Data Sources
Load/Transformation
Business Intelligence
Analytics
Machine Learning
Personalization

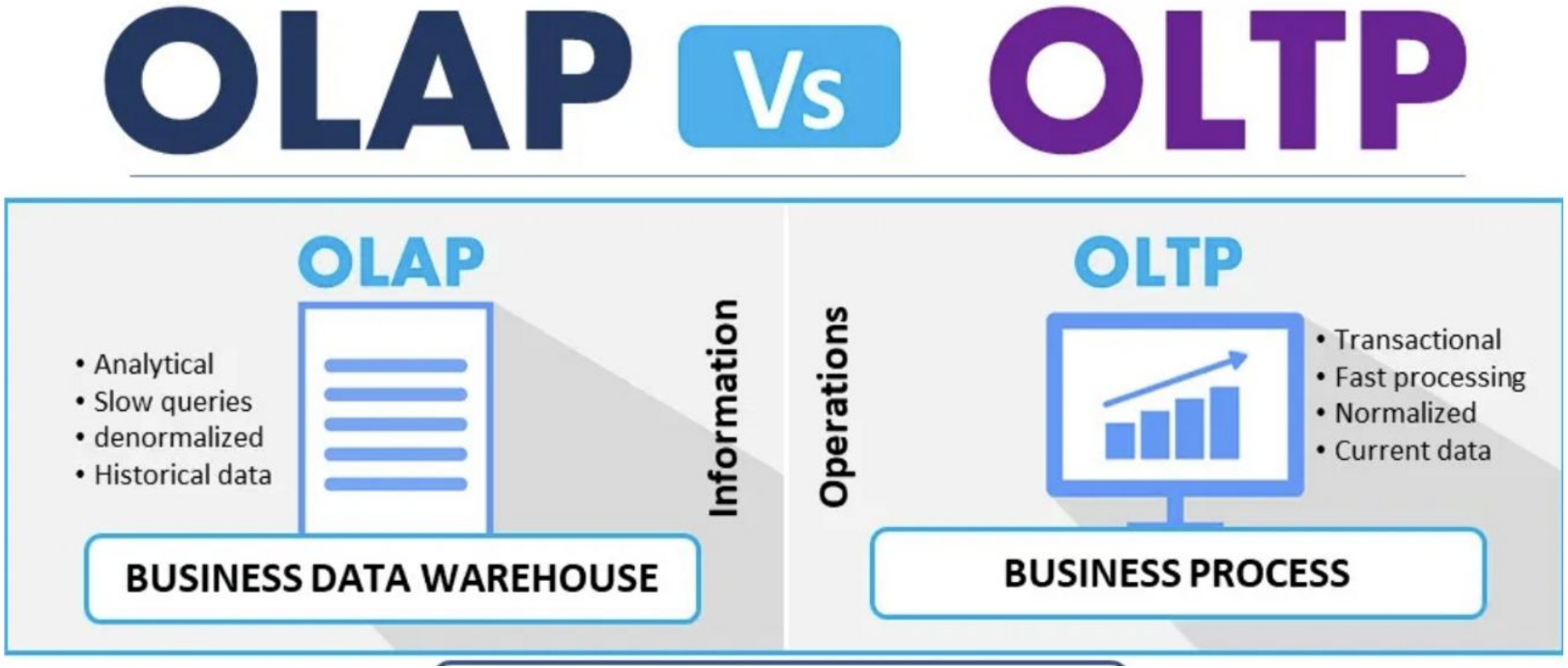
ETL


ETL Process
SQL RDBMS
NoSQL DBMS
SaaS Platforms
XML Files
Extract
Transform
Load
Analyze
Staging Area
Data Warehouse
Analytics

ELT


kafka
Kinesis
CSV, JSON, TXT…
Data Lake
APACHE Spark
Bronze
Raw Ingestion
Silver
Filtered, Cleaned Augmented
Gold
Business-level Aggregates
Streaming Analytics
AI & Reporting

Online Analytical Processing Vs Online Transaction Processing



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

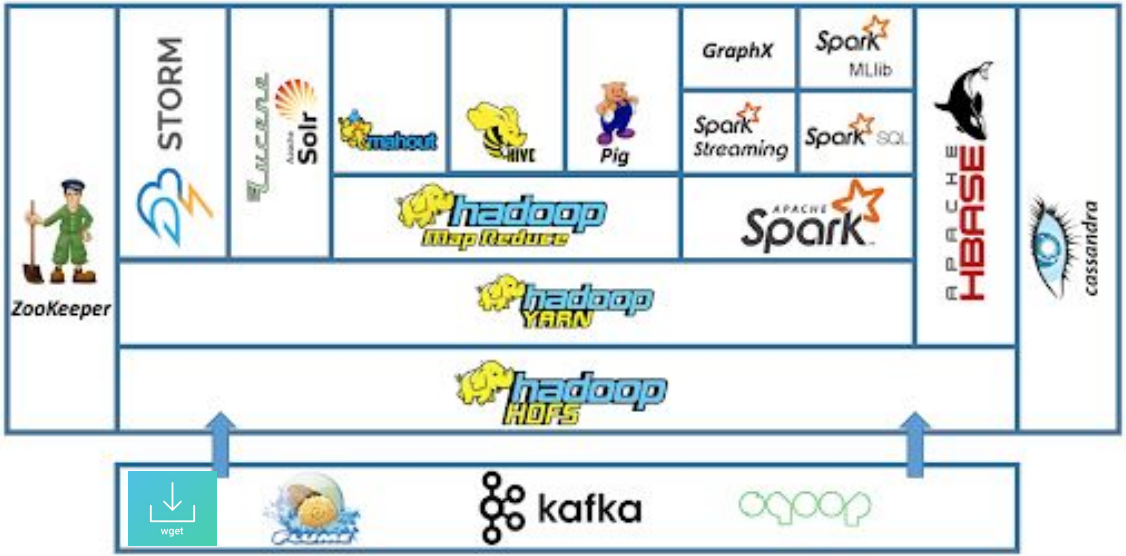
1. Con Sqoop hago un export en archivos parquet/text/avro, etc.
2. Ingesto esos archivos en HDFS (hdfs dfs -put origen destino)
3. Transformo esos archivos con pyspark/scala/sql
4. Cargo esos datos en el Data Warehouse (Hive)

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

- GCP
  - Creación de cuenta y configuración de consola
  - Ingesta de archivos con Gsutil
  - Storage transfer