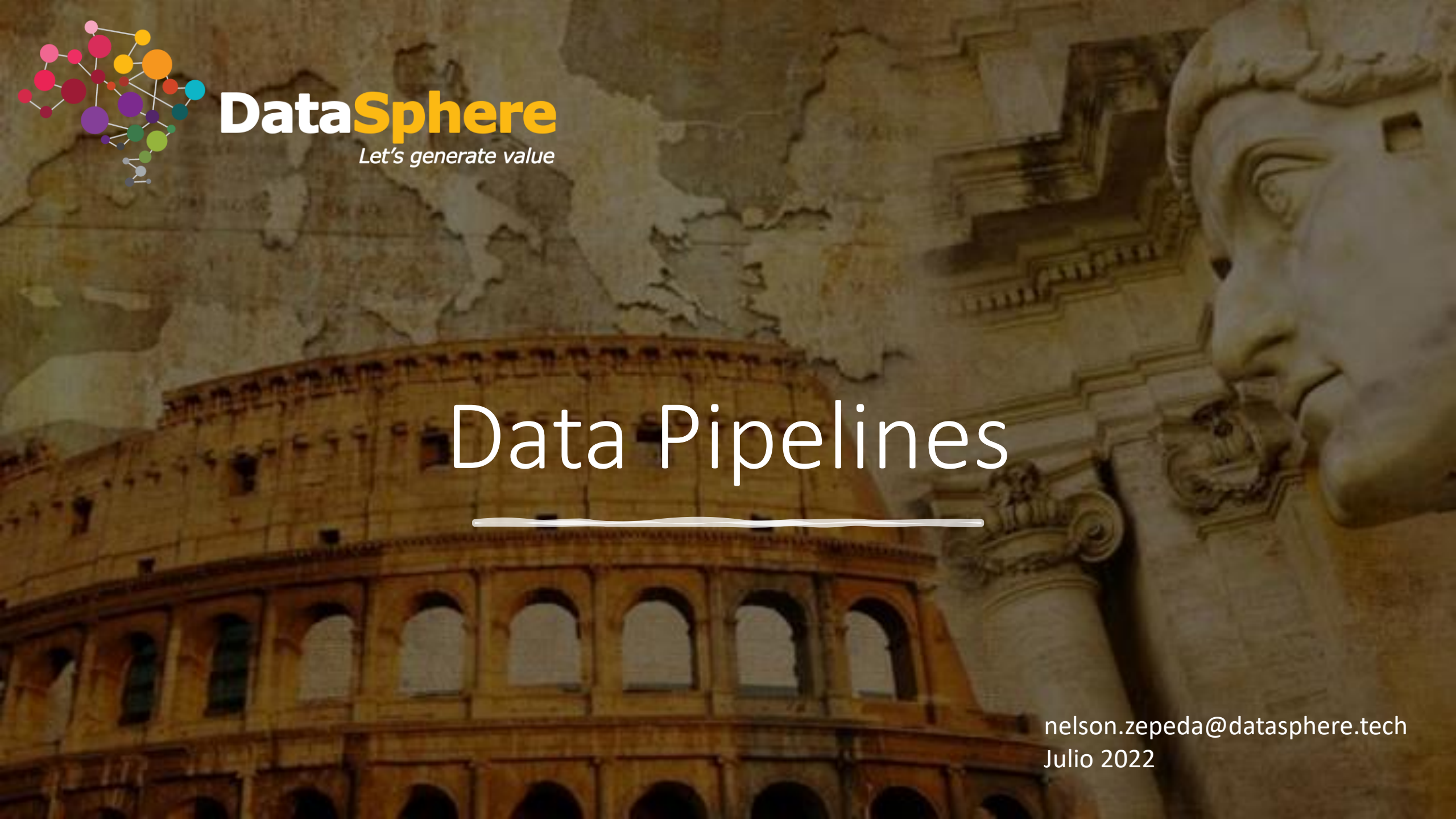
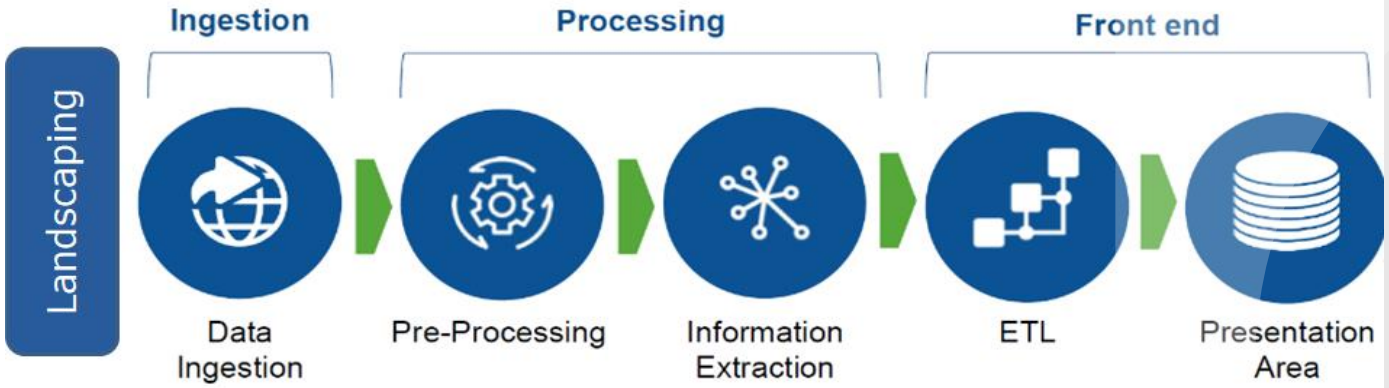



# Data Pipelines

nelson.zepeda@datasphere.tech
Julio 2022

# Acueductos Romanos

# Data pipeline

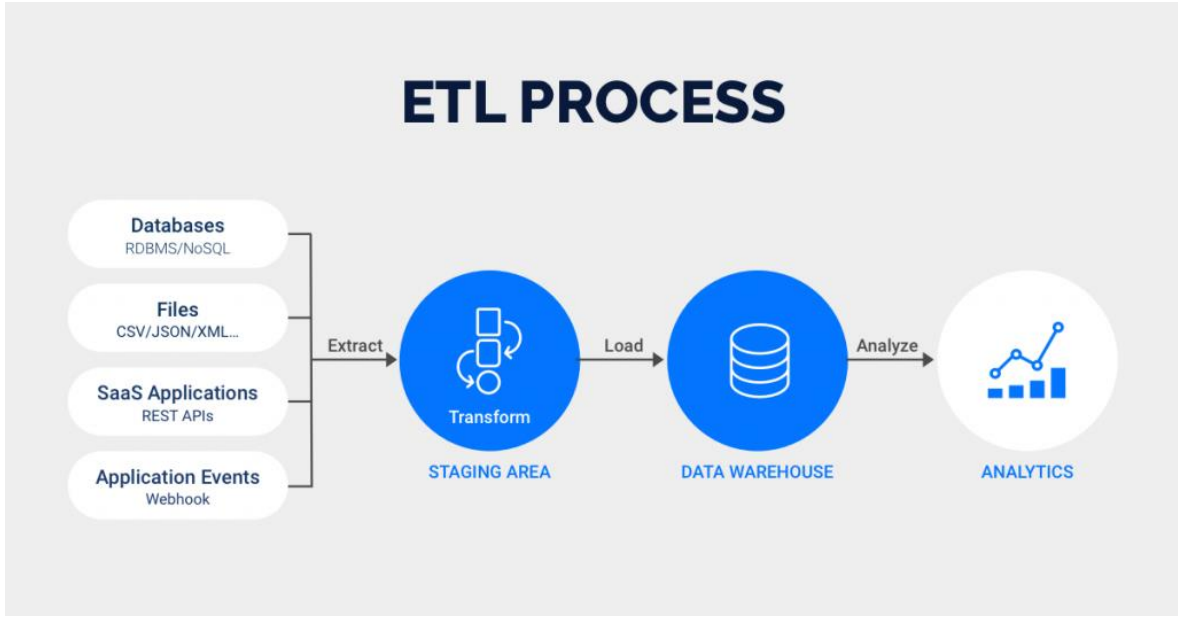



Tomar los datos de numerosas fuentes y procesarlos para proveer un contexto es lo que hace la diferencia entre tener los datos y generar valor con los datos.

Los Data Pipelines son un conjunto de **procesos** que **mueven y transforman** la data de diversas **fuentes** hasta un **destino** capaz de generar valor al ser consumido.

https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Data-pipeline.png

# Data Pipeline: Patrones de Diseño



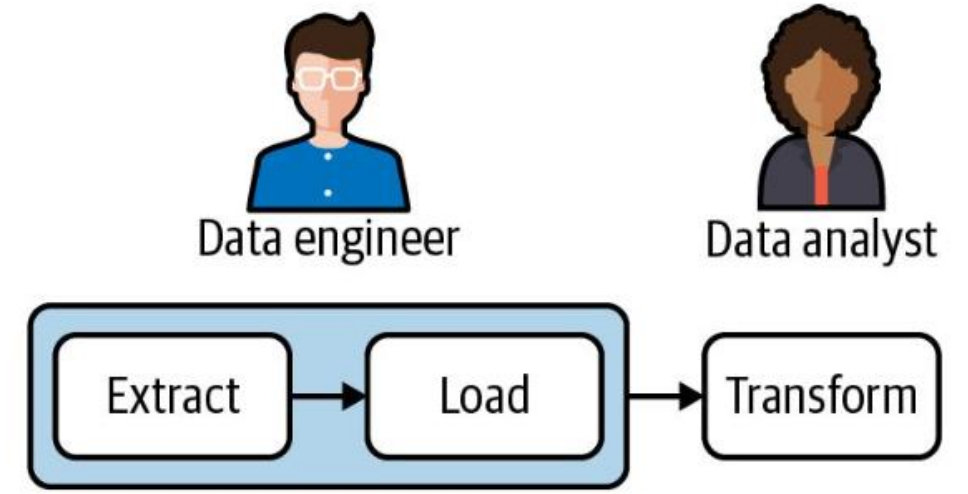
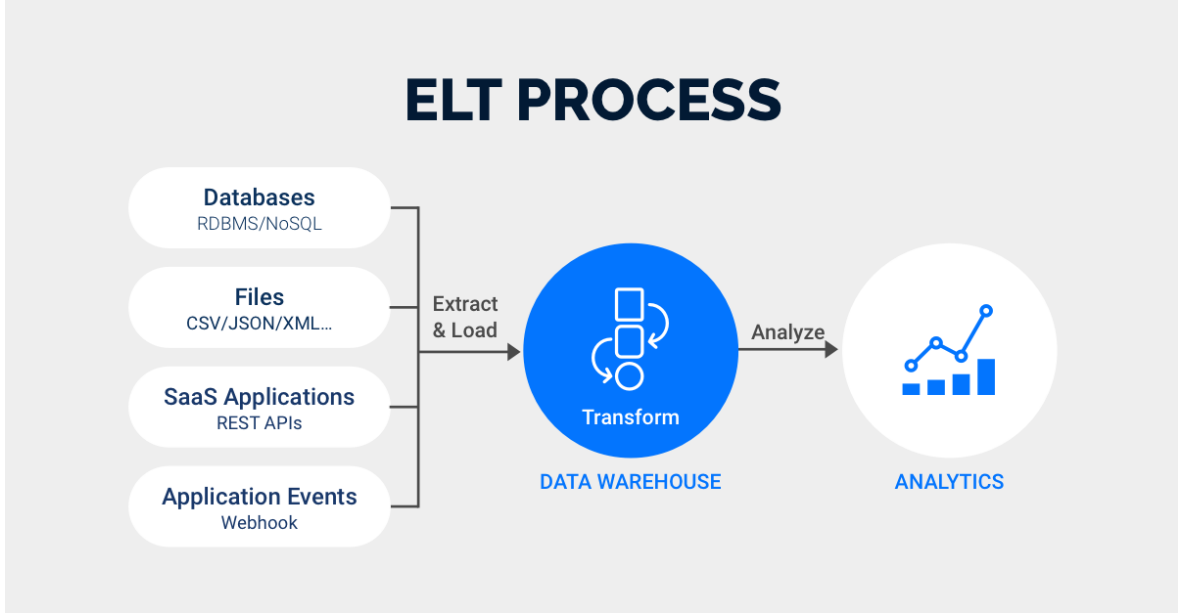




*Figure 3-3. The ELT pattern allows for a clean split of responsibilities between data engineers and data analysts (or data scientists). Each role can work autonomously with the tools and languages they are comfortable in.*




https://rivery.io/blog/etl-vs-elt/

# Data Integration Tools



Source: Gartner (March 2022)

# Caso Práctico

Somos la empresa importadora **Global Products Importer**, el gerente de Mercadeo junto con el gerente de Analytics lograron firmar un acuerdo muy importante con uno de nuestros clientes principales: **El Changarro SA de CV** quienes hace unas semanas implementaron un **proyecto para capturar que tan satisfecho estan los clientes finales con los productos adquiridos** mediante su sitio Web.

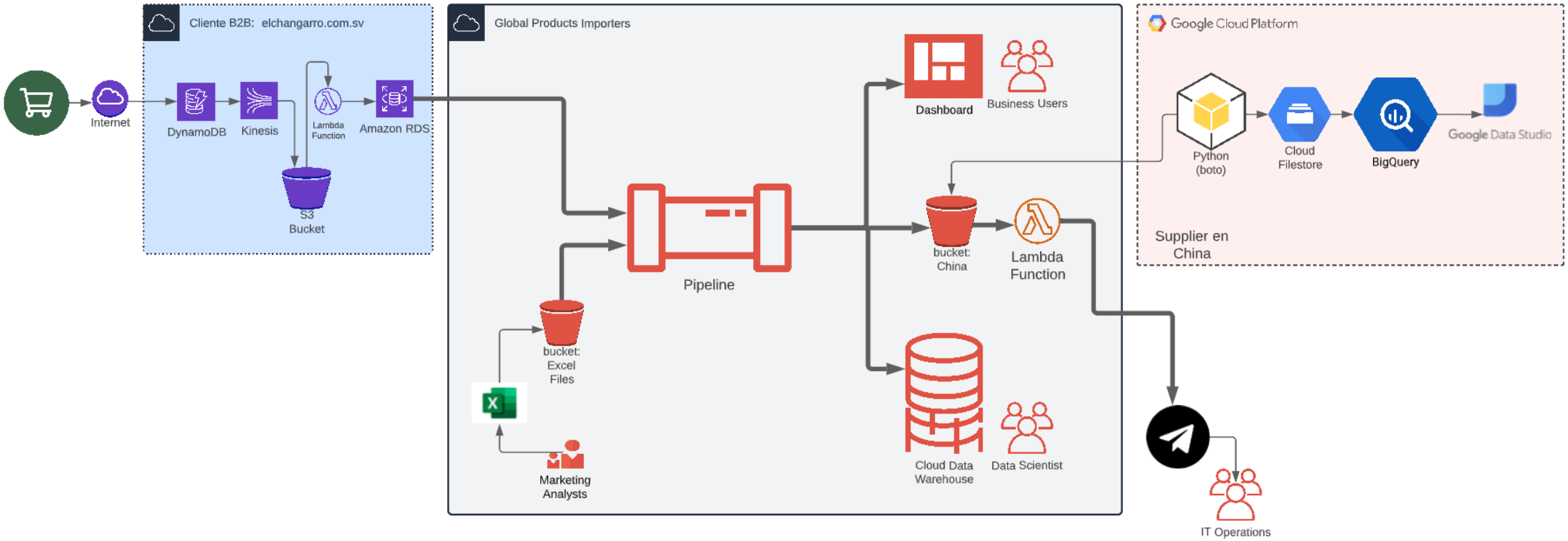
El acuerdo consiste en que nos **compartiran dichos datos** con el fin de que nosotros tengamos visibilidad de como son percibidos los productos que estamos importando desde china. El gerente de mercadeo ha solicitado un **Dashboard** para dar seguimiento y pidio que su equipo de **cientificos de datos** tengan acceso al detalle de los datos cargados.

De forma adicional tambien se solicito incluir **archivos de Excel** que contienen información de los articulos, dichos exceles son generados por un equipo interno de analistas, y como elemento final tambien se pidio **compartir el dataset final con nuestro proveedor en China** quienes procederan a analizar esa data.

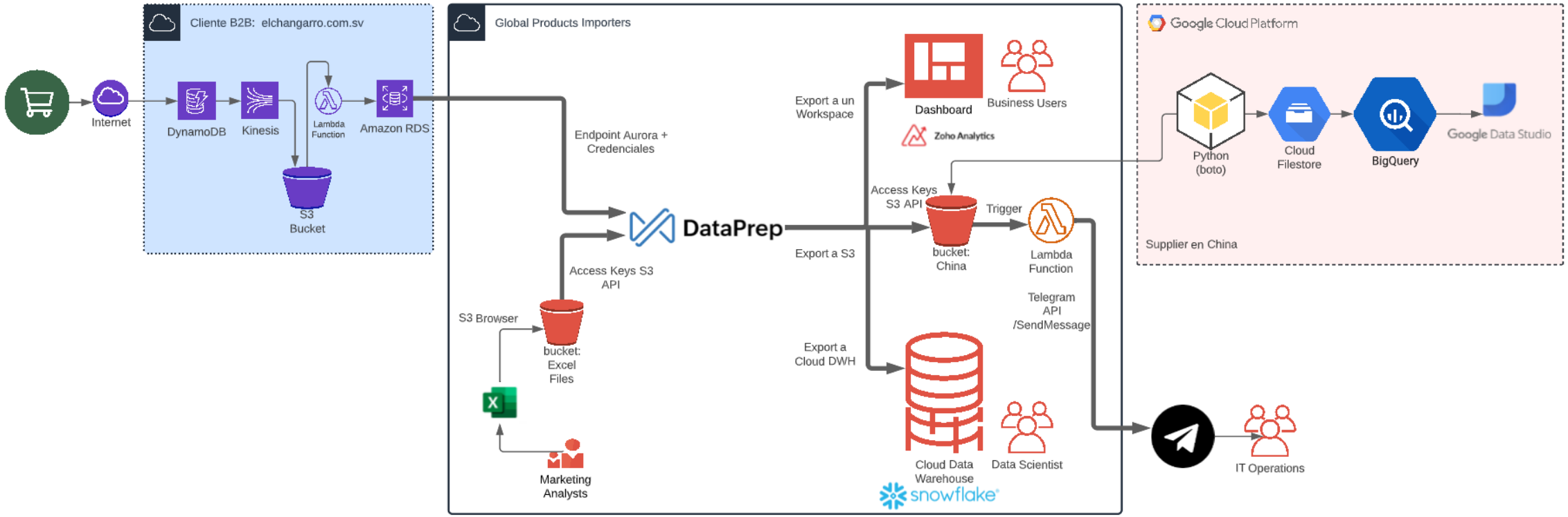
Otros datos:

- El changarro utiliza AWS y su solución es de tipo Serverless pues utiliza DynamoDB, Kinesis, Lambda y RDS, **nosotros tenemos acceso al RDS** que es una base de datos Aurora.
- Nuestros analistas de mercadeo tienen acceso a un repositorio de archivos en S3 y **cargan cada excel utilizando la herramienta S3 Browser**, ellos reemplazan el archivo cada vez que lo actualizan.
- Tenemos un repositorio en **S3 que es exclusivo para nuestro proveedor en china**, ahi le colocamos todos los archivos que necesitamos compartir, ellos utilizan Google Cloud y van a utilizar un script en python para llevarse nuestro archivo y procesarlo para que sea presentado en Google Data Studio una vez los datos esten cargados en su Data Warehouse.
- Dado que la comunicación con China es de alta importancia y somos responsables de que la data se este entregando, la gerencia de operaciones de IT ha pedido que se le notifique mediante un mensaje en **Telegram** cada vez que se deposite un archivo en el repositorio de China
- Los datos se actualizan cada manana (6 am)
- Nosotros utilizarnos AWS y recientemente adquirimos Zoho Analytics y Dataprep

# Diseño


Cliente B2B: elchangarro.com.sv
Internet
DynamoDB
Kinesis
Lambda Function
Amazon RDS
S3 Bucket
Global Products Importers
Pipeline
bucket: Excel Files
Marketing Analysts
Dashboard
Business Users
bucket: China
Lambda Function
Cloud Data Warehouse
Data Scientist
Google Cloud Platform
Python (boto)
Cloud Filestore
BigQuery
Google Data Studio
Supplier en China
IT Operations

# Stack Tecnologías



# Caso Práctico

DataSphere
Let's generate value

# AWS

- RDS: Amazon Relational Database Service (Amazon RDS) es un servicio web que facilita la configuración, la operación y la escala de una base de datos relacional en Nube de AWS.
- S3: Amazon S3 es un servicio de almacenamiento de objetos que almacena datos como objetos dentro de buckets
- S3 Browser: S3 Browser es un muy potente y sencillo cliente de Amazon S3, el servicio de almacenamiento de datos online propiedad de Amazon Web Services.
- Lambda: AWS Lambda es un servicio informático que permite ejecutar código sin aprovisionar ni administrar servidores

# Zoho Analytics

- Zoho Dataprep: Zoho DataPrep ayuda a los científicos de datos con la limpieza de datos a escala, sin ninguna codificación.
- Zoho Analytics: Zoho Analytics es un software de autoservicio de Business Intelligence

# SnowFlake

- Snowflake es una aplicación **SaaS** (*Software as a Service*) basada en el concepto **Data Cloud** (nube de datos);
- Snowflake se puede contratar en tres versiones distintas: **Standard**, **Enterprise** y **Business Critical**, en función de las necesidades del usuario en cuanto a potencia de computación y garantía de disponibilidad.
- Los ámbitos de aplicación donde Snowflake despliega todas sus virtudes son **Data Warehouse** (con su propio motor SQL), **Data Lake**, ingeniería de datos, ciencia de datos, intercambio de datos y desarrollo de aplicaciones de datos.

# Otros Conceptos

- DynamoDB AWS:Amazon DynamoDB es un servicio de base de datos NoSQL
- Amazon Kinesis: Amazon Kinesis le permite incorporar, guardar y procesar datos de streaming en tiempo real
- EPP: Elastic Parallel Processing
- MPP:  Massively Parallel Processing
- SMP (Symmetric Multi-Processing)

# Bibliografía/Referencias


- Data Pipelines Pocket Reference: Moving and Processing Data for Analytics, James Densmore
- Data Pipelines with Apache Airflow, Julian de Ruiter
- https://www.zoho.com/es-xl/analytics/
- https://www.zoho.com/es-xl/dataprep/
- https://aws.amazon.com/es/getting-started/fundamentals-core-concepts/
- https://aprenderbigdata.com/snowflake/