

The background features a soft, pastel color palette with large, organic shapes in shades of pink, peach, and light orange. Delicate white line art of floral branches and leaves is scattered across the design. Small, semi-transparent dots in various pastel colors are also present, adding a textured, bokeh-like effect.

Proyecto Final Análisis de datos

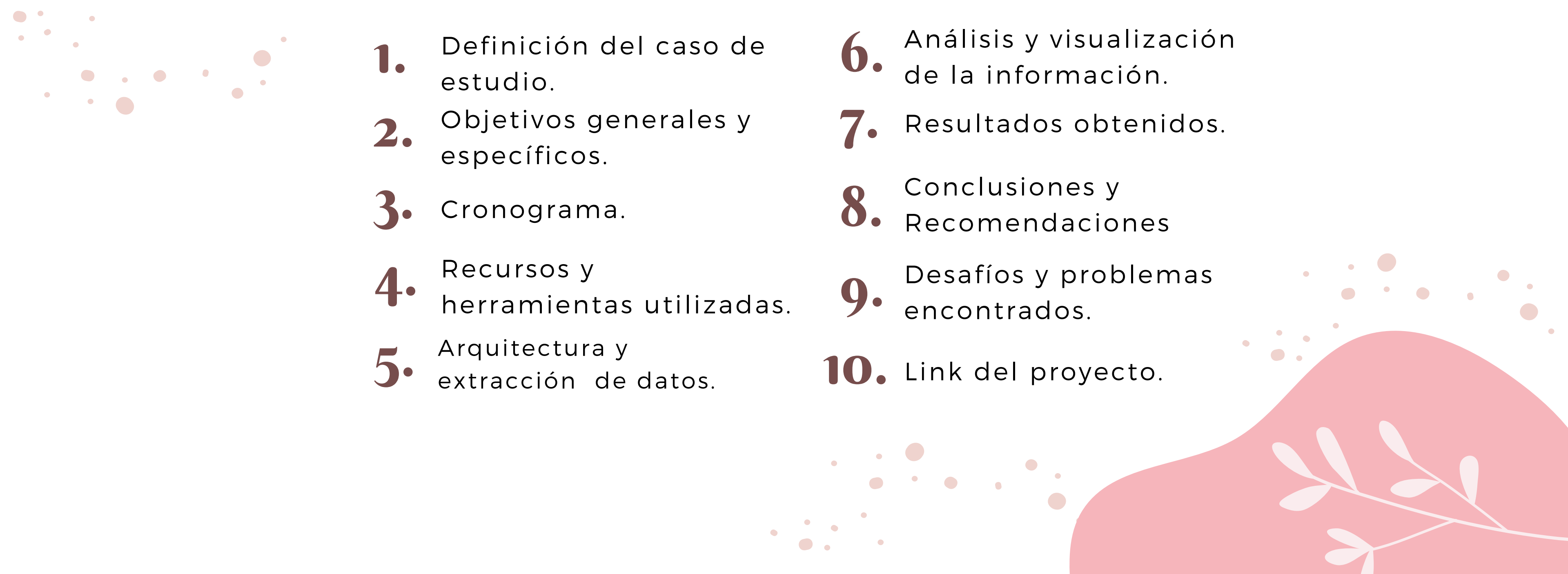
INTEGRANTES:
LUNA SCARLETT
PAZTO ISABEL
MATA JOHN
VELA DAVID

Introducción

Es importante considerar la gran cantidad de datos generados a diario a través de múltiples fuentes, dichos datos son de gran valor si se extraen y analizan correctamente, para lograrlo se propone en el siguiente proyecto diseñar una arquitectura de Data Lake que permita almacenar y procesar grandes volúmenes de datos estructurados y no estructurados provenientes de fuentes como: repositorios públicos, web scraping, archivos csv, json, etc.



Tabla de contenido

- 
1. Definición del caso de estudio.
 2. Objetivos generales y específicos.
 3. Cronograma.
 4. Recursos y herramientas utilizadas.
 5. Arquitectura y extracción de datos.
 6. Análisis y visualización de la información.
 7. Resultados obtenidos.
 8. Conclusiones y Recomendaciones
 9. Desafíos y problemas encontrados.
 10. Link del proyecto.

Temática

1. Definición caso de estudio

Se enfoca en la creación de una arquitectura de Data Lake que permita almacenar y procesar grandes volúmenes de datos estructurados y no estructurados provenientes de diversas fuentes. Para lo cual se usará base de datos SQL y NoSQL, así como Microsoft Azure como concentrador de datos, el cual se conectará con Power BI para realizar el análisis de la información.

2. Objetivo General

Diseñar una arquitectura de Data Lake para la integración de fuentes de datos, su análisis y visualización de indicadores y métricas mediante dashboards en Power BI.

3. Objetivos específicos

Seleccionar al menos 12 fuentes de datos a integrar en la solución respecto a los temas propuestos.

Realizar la limpieza de datos respectiva a cada fuente.





Definir cinco casos de estudio teniendo en cuenta los temas seleccionados y la coherencia de los datasets.

Concentrar todos los datos en un repositorio centralizado.

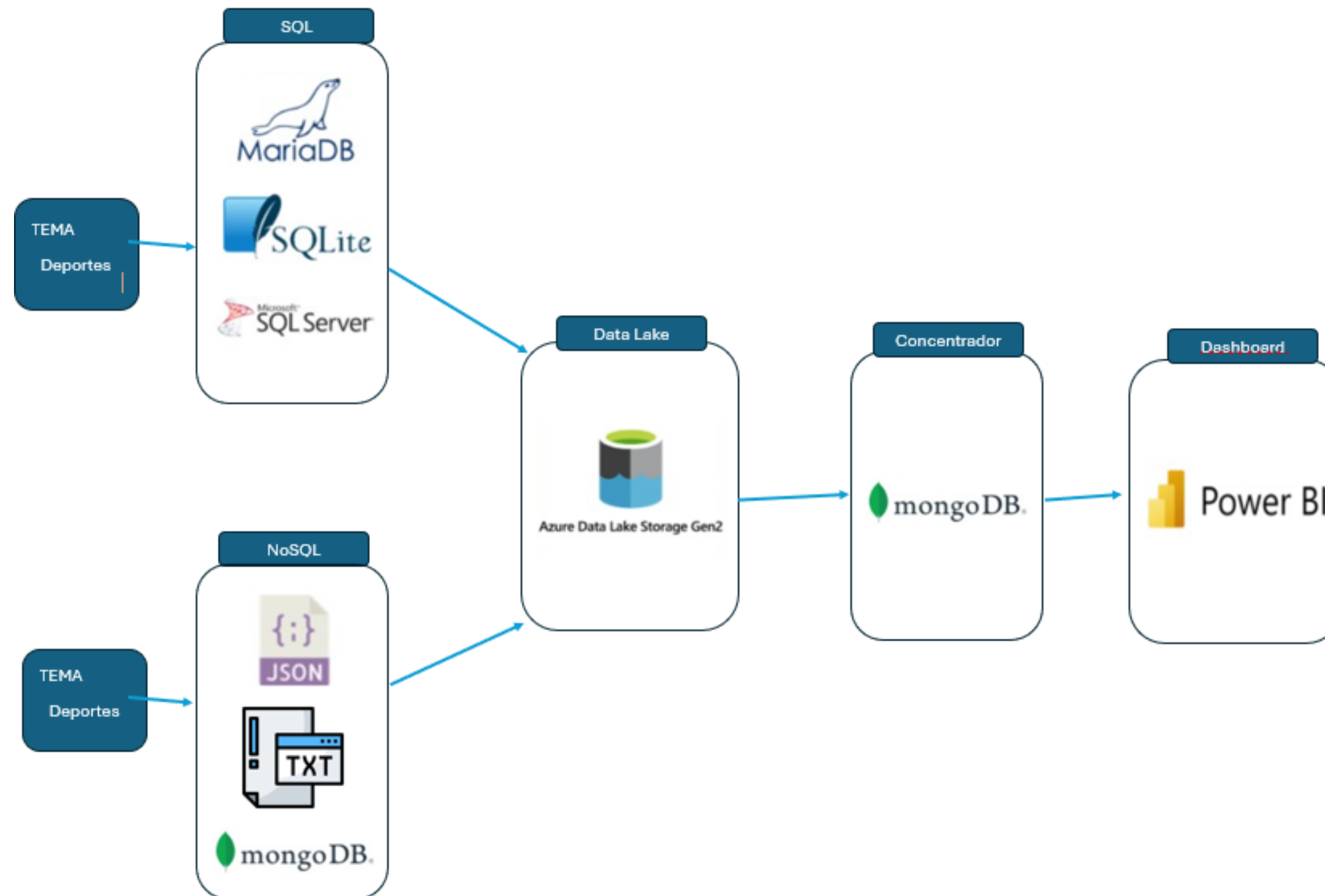
Identificar los índices y métricas a evaluar para cada caso de estudio conforme a los datos que le corresponden.

Generar el dashboard correspondiente a cada caso de estudio haciendo uso de Power BI.

Punto #4 Recursos y herramientas utilizadas

Recursos y herramientas utilizadas	
Fuentes de los datos: Repositorios públicos, web scraping, archivos csv, json, entre otros.	
Base de datos SQL: MySQL, Azure Data Lake Gen2, SQL Server, PostgreSQL y MongoDB.	
Lenguaje de programación: Python.	
Análisis de datos: Power BI.	

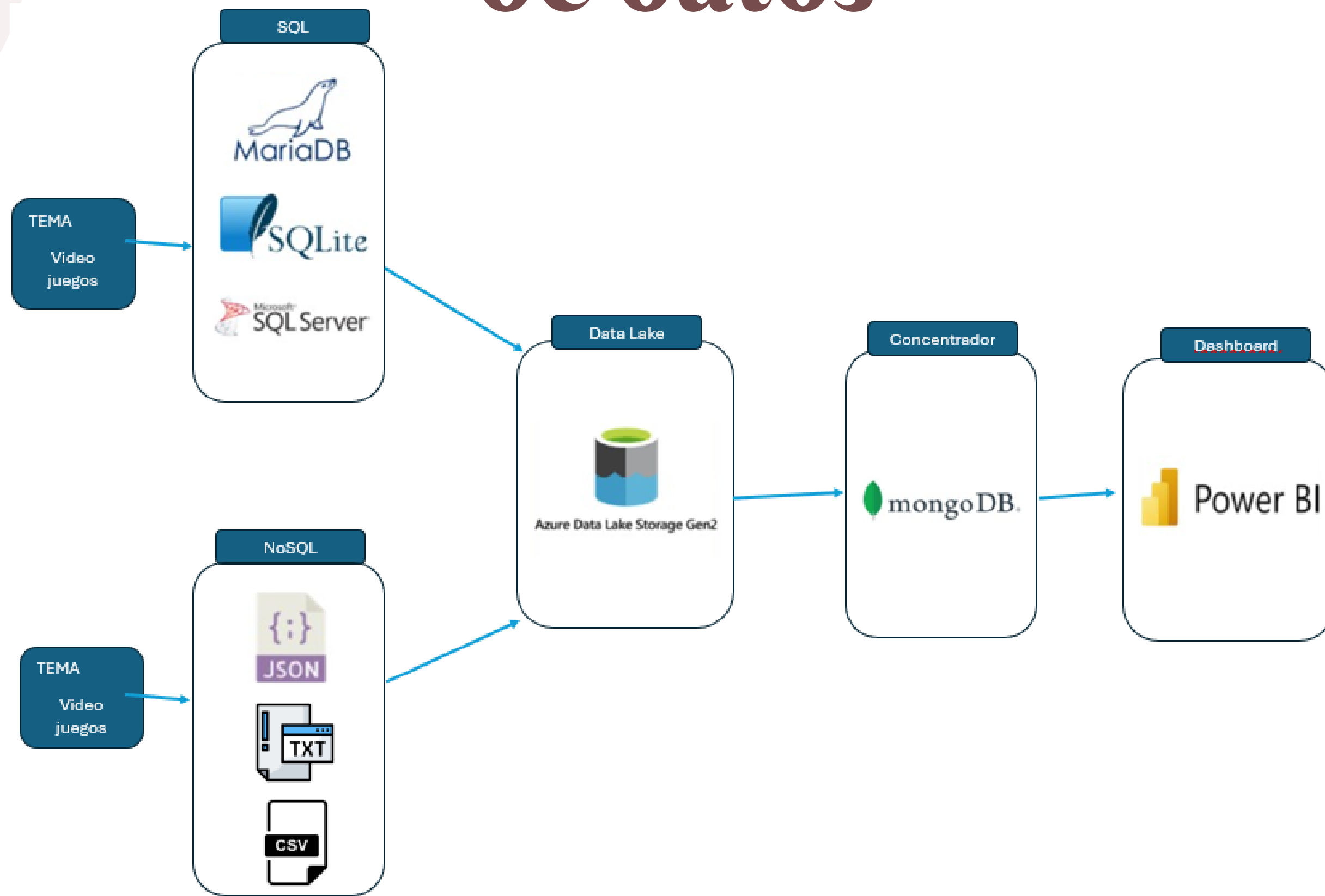
Punto #5 Arquitectura de datos



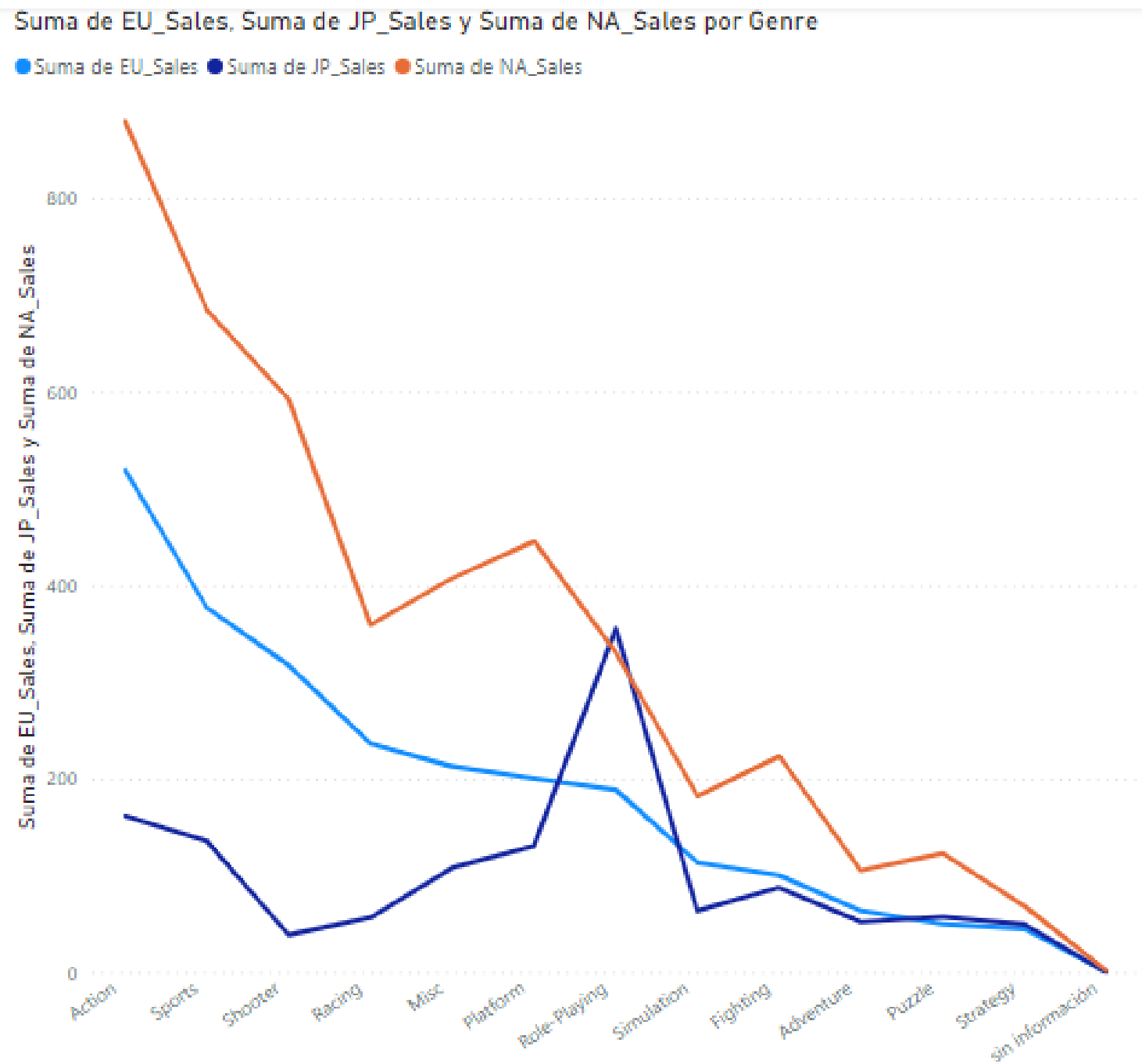
Suma de match_id por team



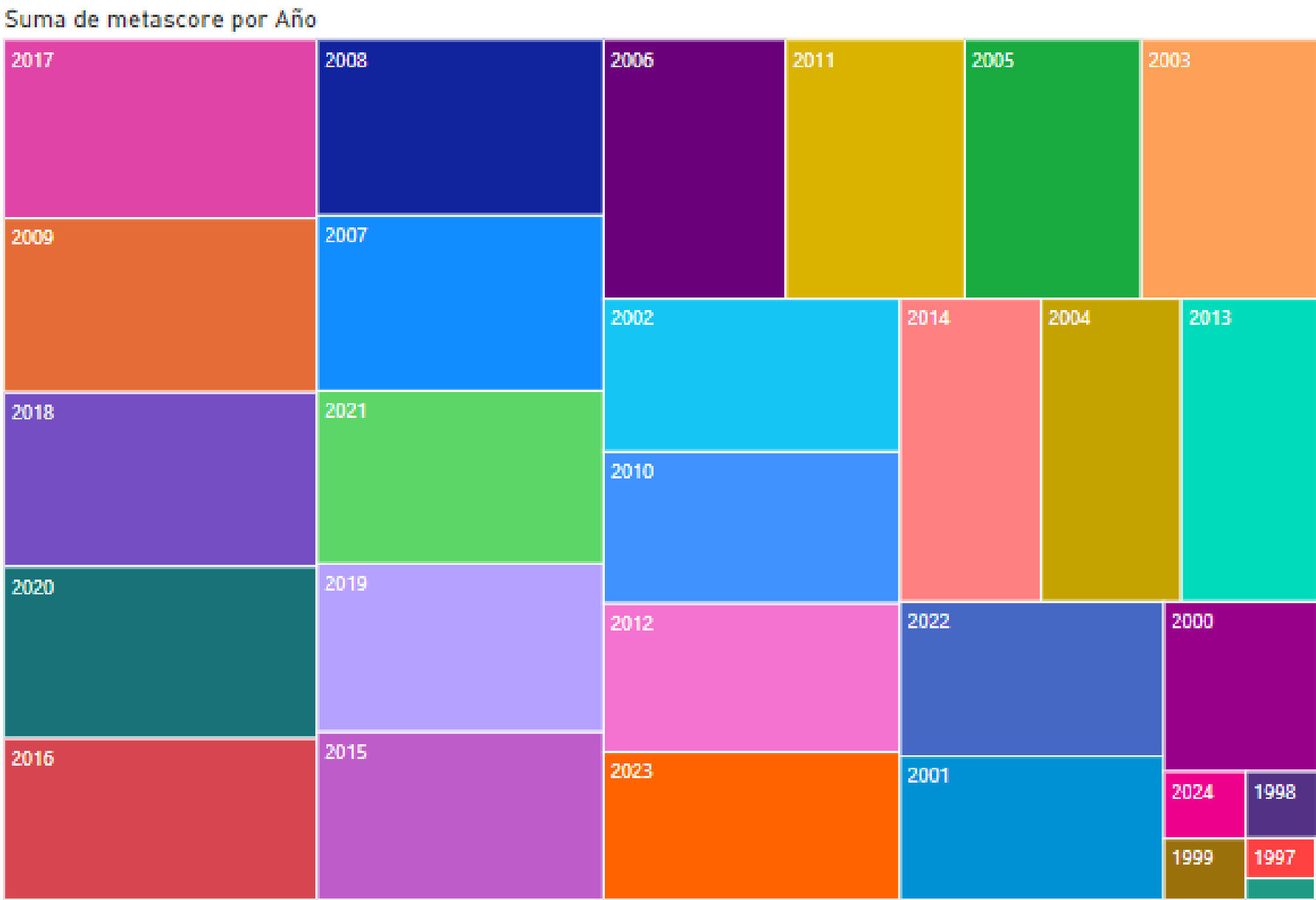
Punto #5 Arquitectura de datos



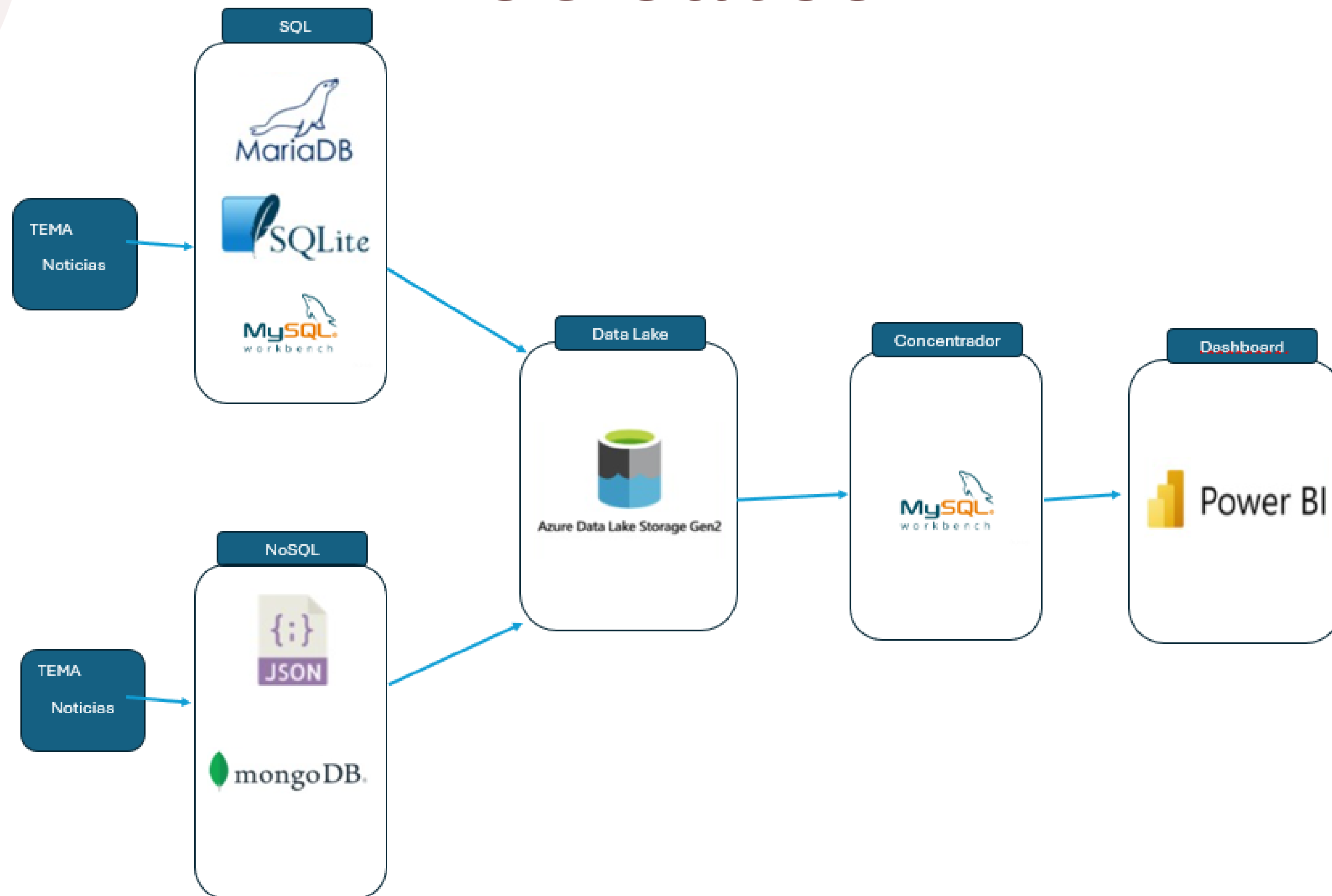
Generos mas vendidos por región



Años con mejores puntuaciones generales

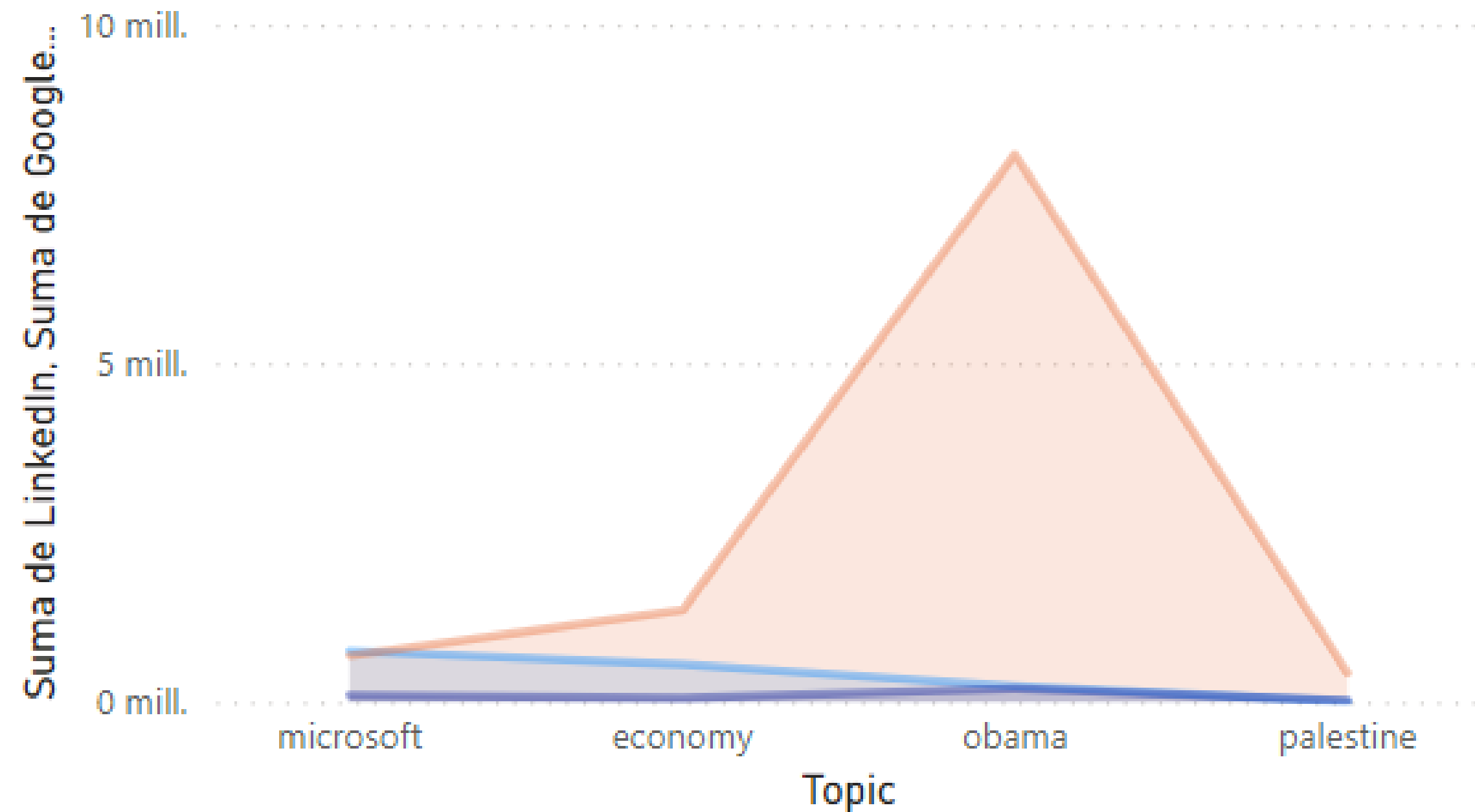


Punto #5 Arquitectura de datos



Suma de LinkedIn, Suma de GooglePlus y Suma de Facebook por Topic

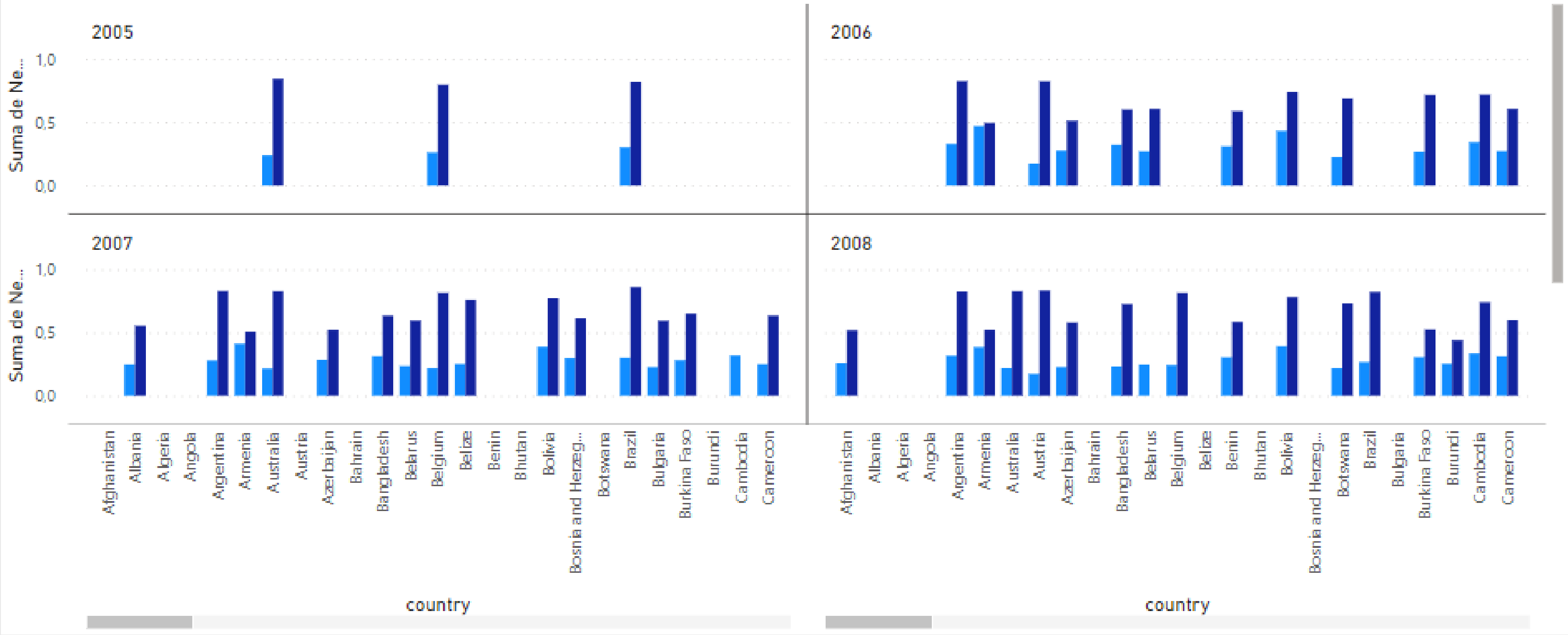
● Suma de LinkedIn ● Suma de GooglePlus ● Suma de Facebook



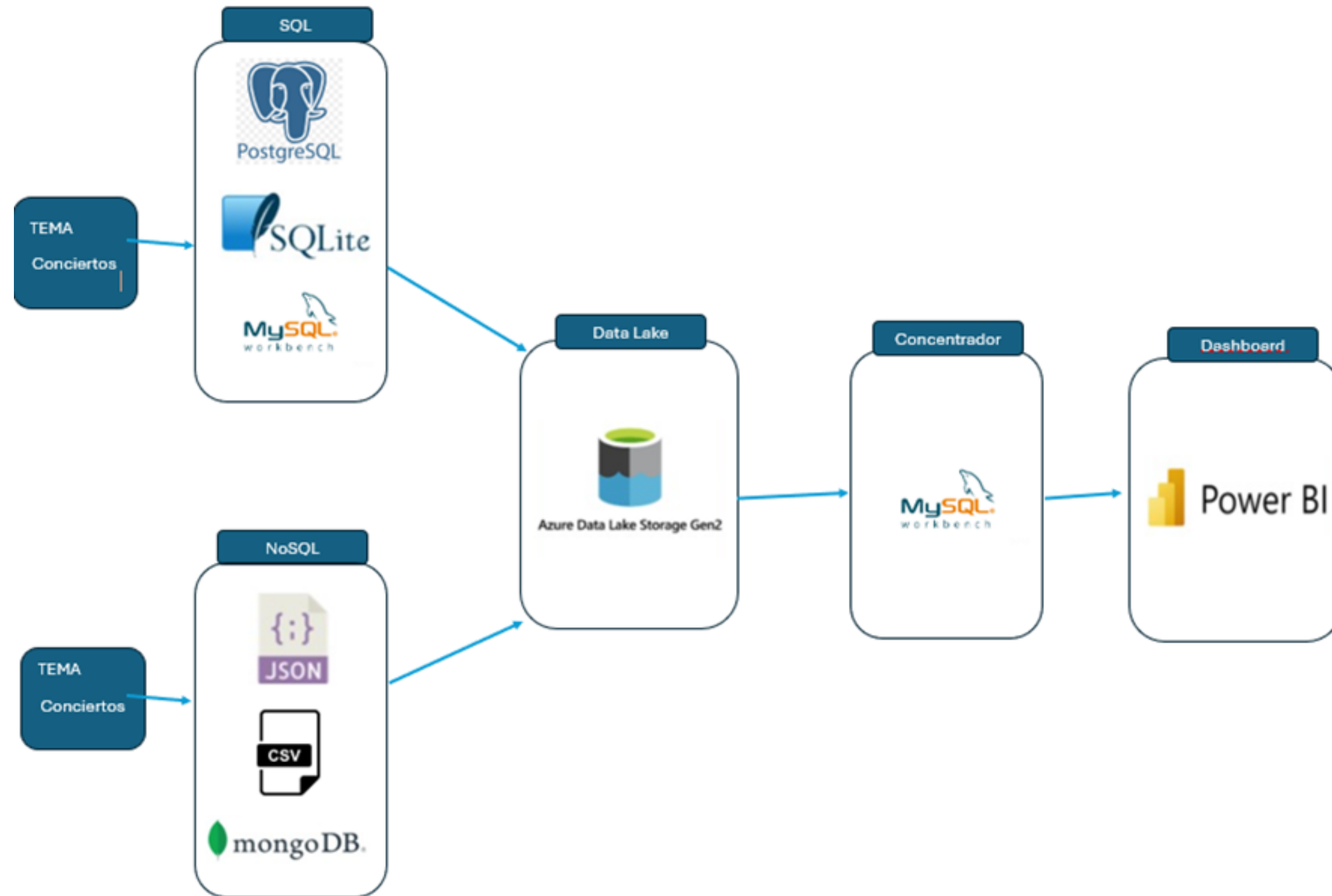
Suma de efecto negativo y suma de efecto positivo por pais y año

Suma de Negative affect y Suma de Positive affect por country y year

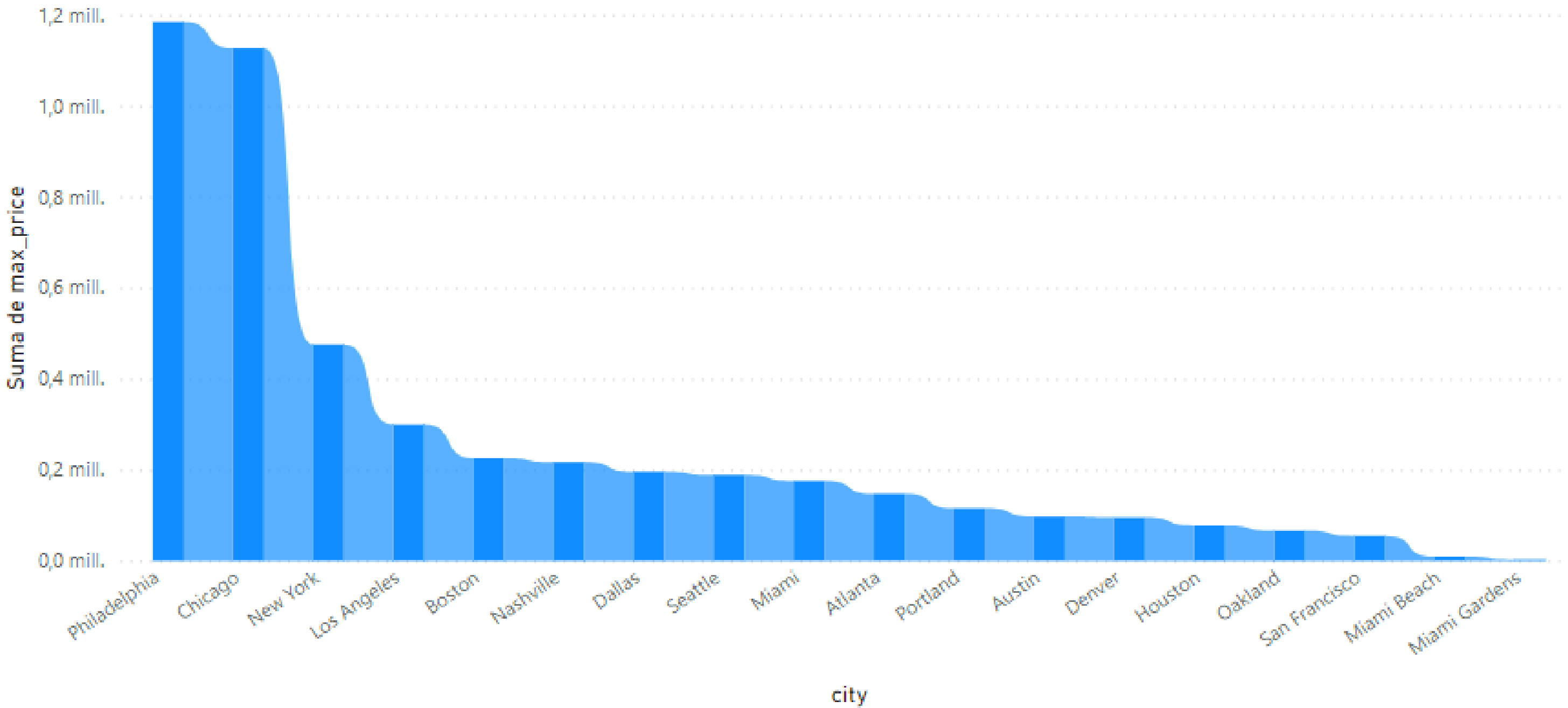
Suma de Negative affect Suma de Positive affect



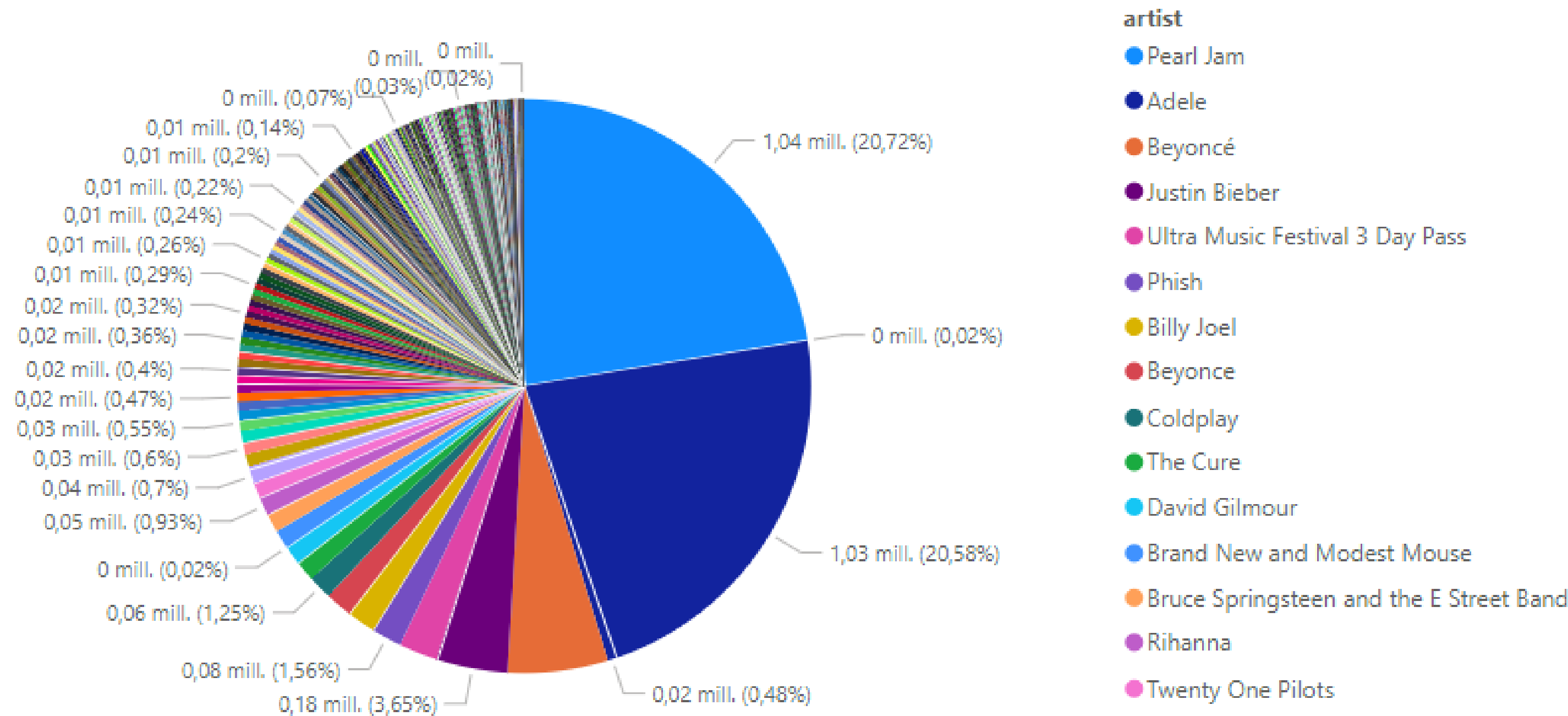
Punto #5 Arquitectura de datos



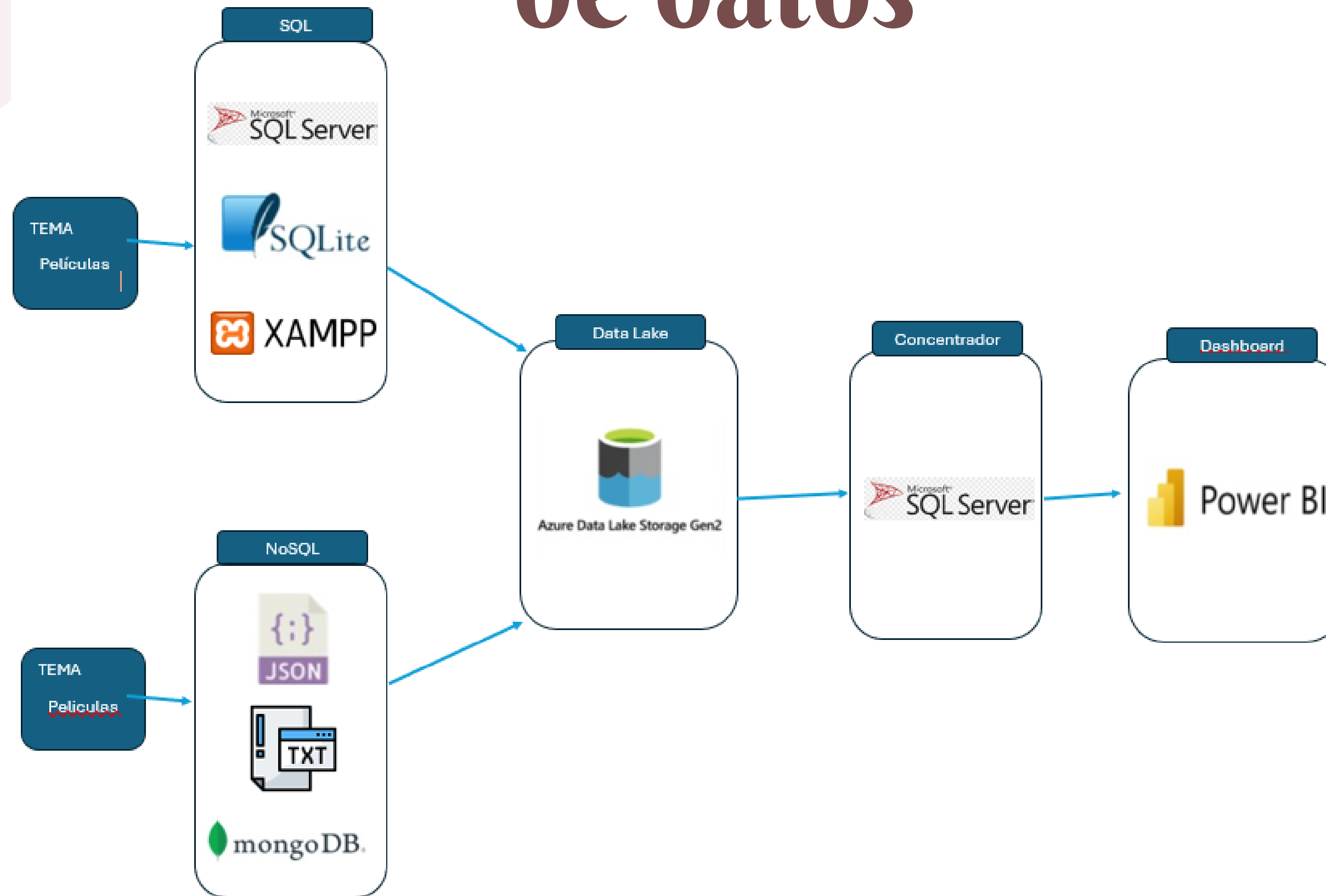
Suma de max_price por city



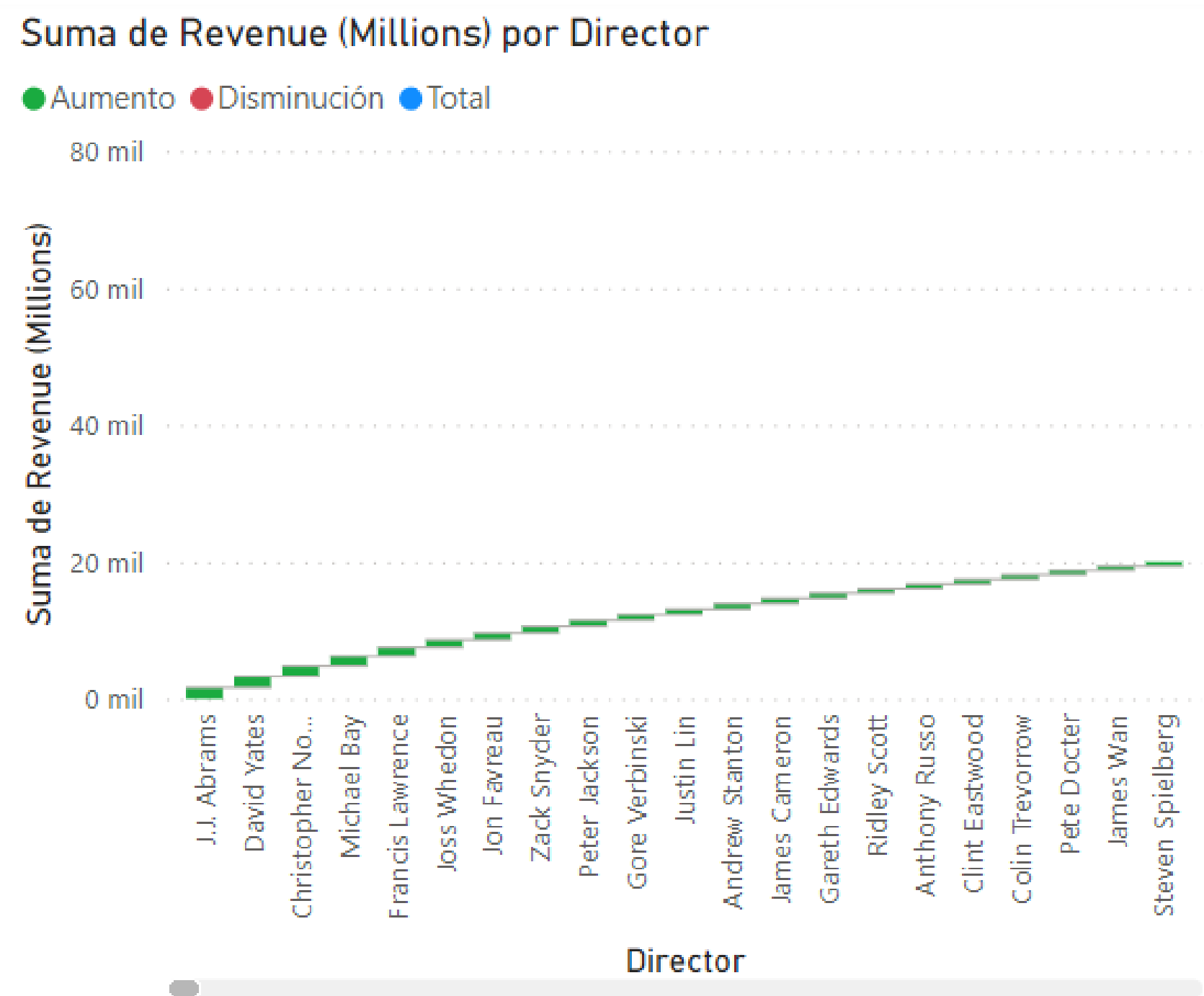
Suma de max_price y Suma de min_price por artist



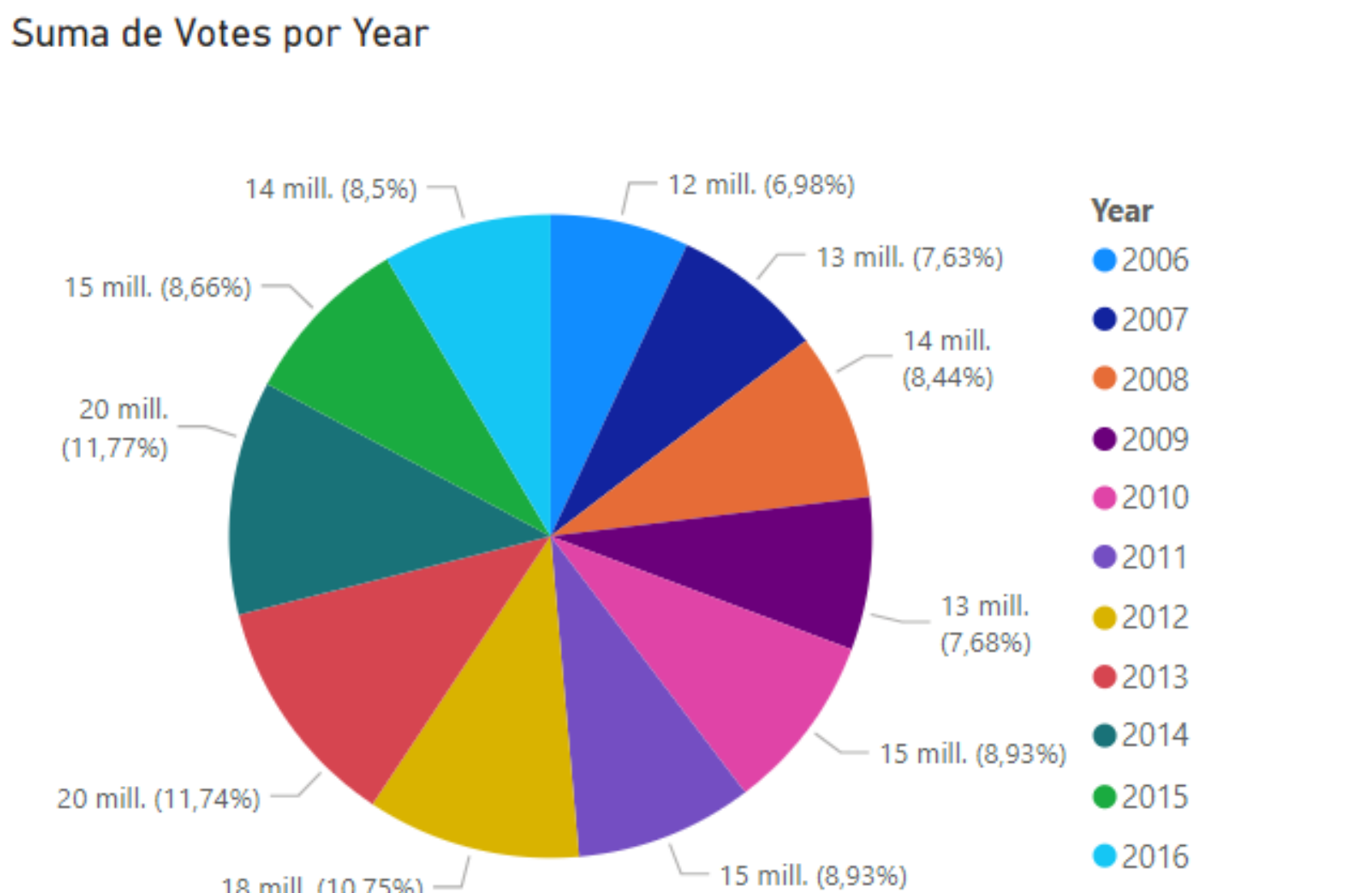
Punto #5 Arquitectura de datos



Directores con mayores remuneraciones



Numero de votantes de peliculas por año



Extracción de datos

Proceso de adquisición de datos

Las fuentes que se usaron para encontrar los datasets fueron:

- Worldbank
- Google Public Data
- Kaggle
- Datasetsearch
- Public tableau
- Census
- Scrapy

Conclusiones

- Se implementó una arquitectura de Data Lake que permitió la recolección, integración y almacenamiento de grandes volúmenes de datos.
- La arquitectura de Data Lake utiliza Microsoft Azure como plataforma en la nube, integrando con las bases de datos.
- Se identificaron y analizaron los índices y métricas relevantes para cada uno de los casos de estudio.
- Se realizaron los dashboards respectivos.

Recomendaciones

- Realizar una buena investigación para la recolección de datasets sea eficiente, tenga calidad y sea veraz.
- Automatizar en la medida de lo posible los procesos de recolección, limpieza y carga de datos en el Data Lake con el fin de optimizar el flujo de trabajo.

Desafío

- El desafío principal al momento de realizar este proyecto fue la búsqueda de datasets, En algunos casos, fue difícil encontrar datasets que se ajustaran a las necesidades específicas del proyecto, lo que requirió un esfuerzo adicional por parte del equipo para buscar alternativas y adaptar los datos disponibles.

Links

- <https://drive.google.com/file/d/1SpNsZ7fbf-KET4xFkJ7inhcdvoFTJ1XY/view?usp=sharing>

