Federico Gonzalez Cardenas
Isaac Piedrahita Carvajal

Javier Alejandro Vergara Zorrilla
ETL

Ingeniería de Datos e Inteligencia Artificial, Facultad de ingeniería
Universidad Autónoma de Occidente
Santiago de Cali
2024

Aspectos a tener en cuenta:

- El proceso de extracción y consulta se realizó con dos APIs (Financial Modeling y Yahoo Finance) buscando así poder trabajar con un dataset muchísimo más completo al que solo una API nos podía ofrecer con datos y columnas muy limitadas.

- La carpeta "API" contiene el archivo python que hace referencia únicamente al codigo de extracción que le hicimos, también esta carpeta contiene dos EDAs los cuales tienen el mismo proceso con la diferencia que el notebook API_ETL_EDA se trabajo con un csv el cual contiene todo el dataframe obtenido del proceso de extracción, en ese EDA hacemos todo el proceso desde ese csv para poder optimizar el trabajo a lo largo de la realización del proyecto. Por otra parte también tenemos el notebook API_ETL_EDA_NOT_CSV el cual hace el mismo análisis que el EDA original pero extrayendo directamente los datos del data frame que obtuvimos gracias a las APIs.
  (Todo esto con el fin de mostrar que efectivamente el grupo de trabajo cumple con realizar el EDA solo obteniendo los datos del proceso de extracción conectando y consultando las APIs, pero también dejando en claro la opción que decidimos tomar de CSV para poder trabajar el EDA sin tener que repetir el proceso de extracción cada rato a lo largo de este mes).

- Se realizaron dos dashboards, uno del modelo dimensional enfocado a la tendencia que encontramos en el primer corte el cual era que más de un 60% de los registros de los despidos del dataset eran de estados unidos, pudiendo entregar gracias a la realización de un modelo dimensional un dataset enfocado solo a los despidos de la industria en estados unidos.

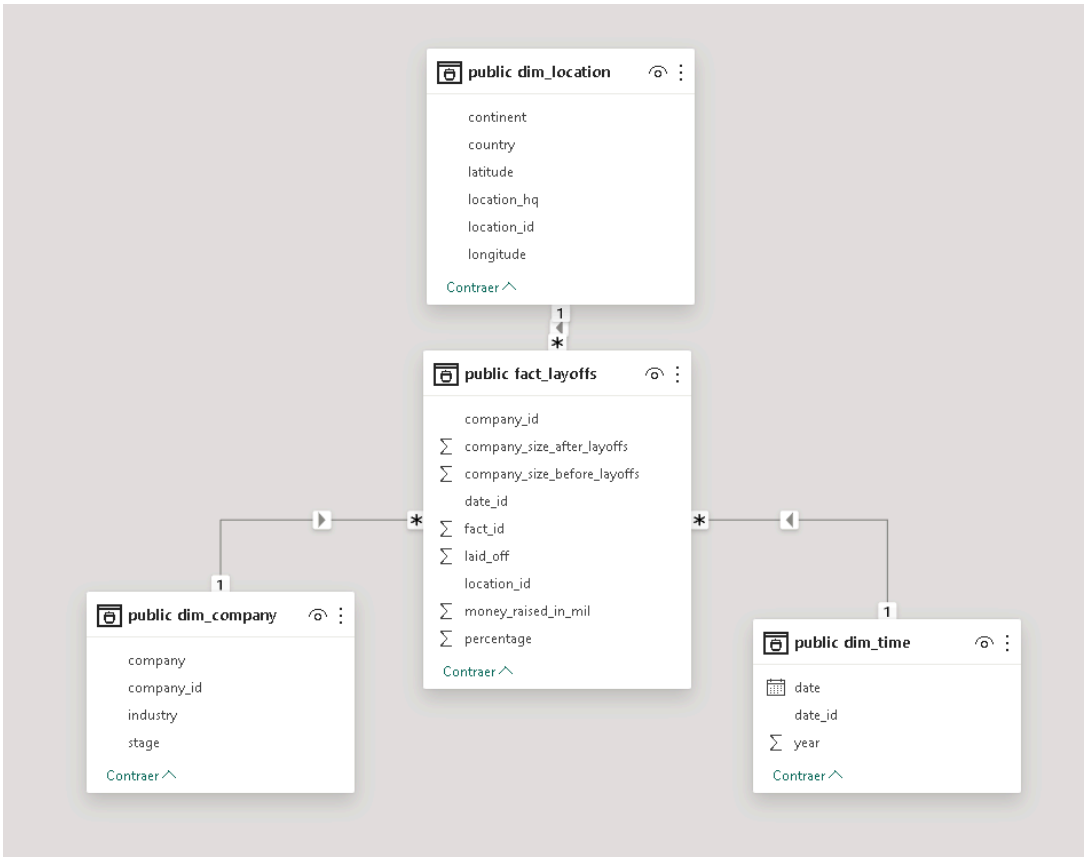
- Cabe recalcar que en el dashboard del modelo dimensional no se tuvieron en cuenta los años 2021 y 2024 ya que estos desde el primer análisis del proyecto vimos que tenían demasiados pocos registros, y queriendo seguir cada vez un enfoque más concreto y de utilidad en nuestro análisis decidimos eliminar esos años en nuestro dashboard para solo trabajar con años que aporten un verdadero valor a nuestras gráficas.

- El otro dashboard está enfocado solamente al dataset que obtuvimos como producto final de nuestro EDA.

- La carpeta "Data_API" contiene el csv dataset_api, el cual es el csv que extrajimos de todo el proceso de consultas que hicimos con las dos APIs (este archivo es el que usa el eda llamado "API_ETL_EDA"). También esta carpeta tiene otro csv llamado stock_data el cual es el data frame que únicamente sacabamos solo consultando la api de financial modeling, ese csv lo dejamos con el fin de mostrar cómo de alto era el valor que le añadimos al dataset usando la otra API al final.

Evidencias:

- Diagrama modelo dimensional (obtenido al importar las dimensiones en Power BI)



- Proceso de extracción de los datos de la API (Este proceso también lo encontrarás en el archivo API_ETL_EDA_NOT_CSV)

```
Procesando 1/57525: NDBKY
Procesando 2/57525: NRRWF
Procesando 3/57525: 5309.KL
Procesando 4/57525: TQGEX
Procesando 5/57525: MEA.AX
Procesando 6/57525: BNP.PA
Procesando 7/57525: SLQT
Procesando 8/57525: 0298.HK
Procesando 9/57525: RAFL.BO
Procesando 10/57525: ANDE
Procesando 11/57525: KITL
Procesando 12/57525: BYOT.L
Procesando 13/57525: 001065.KS
Procesando 14/57525: JVTNX
Procesando 15/57525: VSL.NZ
Procesando 16/57525: 7076.KL
Procesando 17/57525: 0613.HK
Procesando 18/57525: 002732.SZ
Procesando 19/57525: G0VB
```

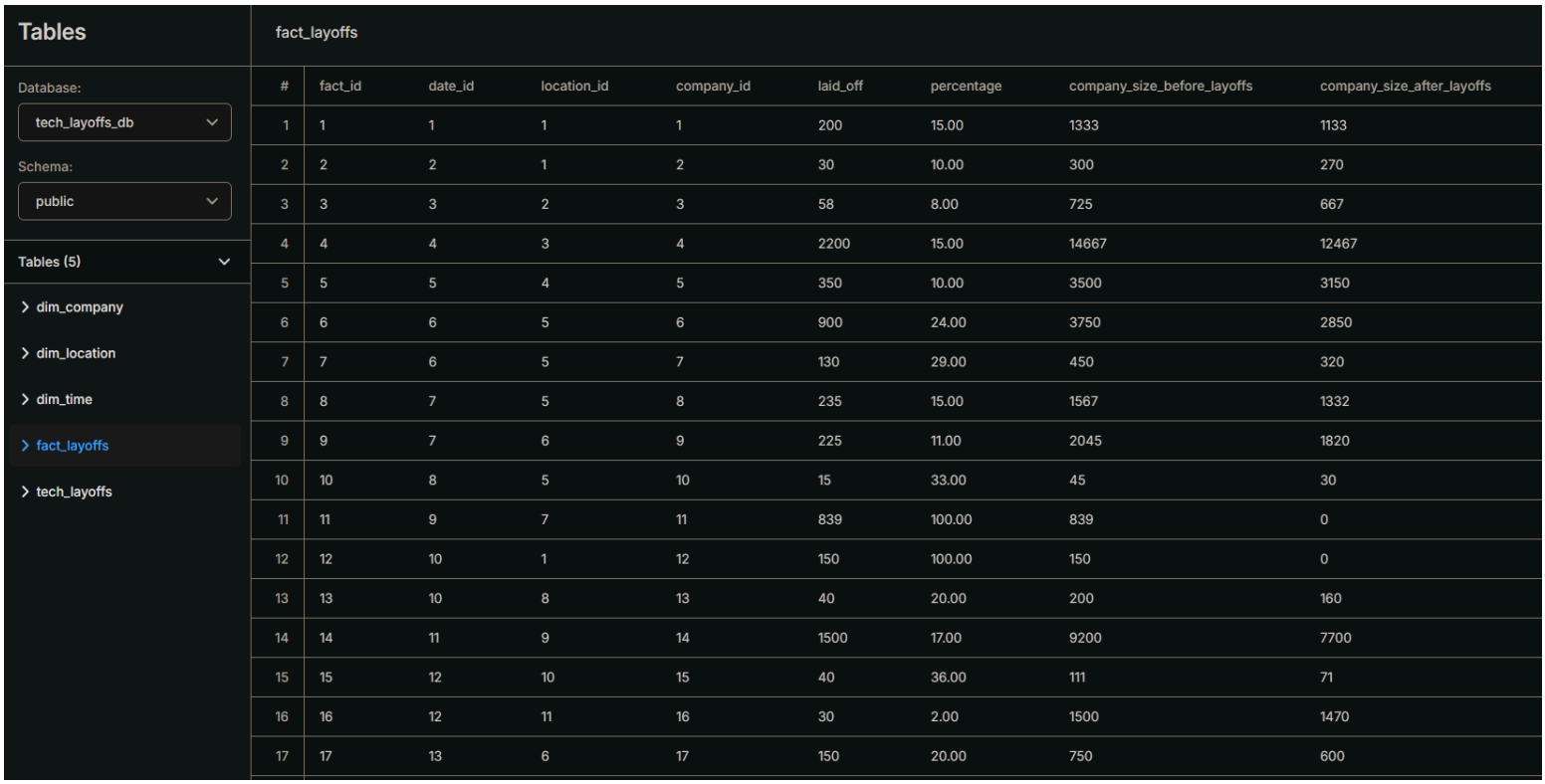
```
Procesando 57524/57525: 1727.HK
Procesando 57525/57525: 300621.SZ
    symbol                                name    price  \
0    NDBKY                Nedbank Group Limited  11.4900
1    NRRWF                  Nuran Wireless Inc.   0.0909
2  5309.KL                  ITMAX System Berhad   2.1900
3    TQGEX  T. Rowe Price QM Global Equity Fund  16.7800
4   MEA.AX                       McGrath Limited   0.5950

                       exchange exchangeShortName  \
0                    Other OTC               PNK
1                    Other OTC               PNK
2                  Kuala Lumpur               KLS
3                        NASDAQ            NASDAQ
4  Australian Securities Exchange               ASX

                       Industry Profit Margins       PE Ratio  \
0              Banks - Regional        0.25108       6.921687
1       Communication Equipment  No disponible  No disponible
2  Security & Protection Services        0.42276           36.5
```

Recordemos que el proceso que se sigue es que gracias al carácter “symbol” que nos da el primer dataset de la API de financial modeling, es que gracias a la biblioteca de la API de Yahoo Finance hacemos una consulta de información para cada symbol que nos dieron al principio.

Tablas del modelo dimensional (Base de datos **PSQL** serverless)



Tables

Database: tech_layoffs_db

Schema: public

Tables (5)

- dim_company
- dim_location
- dim_time
- fact_layoffs
- tech_layoffs

fact_layoffs

#	fact_id	date_id	location_id	company_id	laid_off	percentage	company_size_before_layoffs	company_size_after_layoffs
1	1	1	1	1	200	15.00	1333	1133
2	2	2	1	2	30	10.00	300	270
3	3	3	2	3	58	8.00	725	667
4	4	4	3	4	2200	15.00	14667	12467
5	5	5	4	5	350	10.00	3500	3150
6	6	6	5	6	900	24.00	3750	2850
7	7	6	5	7	130	29.00	450	320
8	8	7	5	8	235	15.00	1567	1332
9	9	7	6	9	225	11.00	2045	1820
10	10	8	5	10	15	33.00	45	30
11	11	9	7	11	839	100.00	839	0
12	12	10	1	12	150	100.00	150	0
13	13	10	8	13	40	20.00	200	160
14	14	11	9	14	1500	17.00	9200	7700
15	15	12	10	15	40	36.00	111	71
16	16	12	11	16	30	2.00	1500	1470
17	17	13	6	17	150	20.00	750	600

Creación tabla de hechos y modelos:

```sql
CREATE TABLE fact_layoffs (
    fact_id SERIAL PRIMARY KEY,
    date_id INT,
    location_id INT,
    company_id INT,
    laid_off INT,
    percentage DECIMAL(5,2),
    company_size_before_layoffs INT,
    company_size_after_layoffs INT,
    money_raised_in_mil DECIMAL(12,2),
    FOREIGN KEY (date_id) REFERENCES dim_time (date_id),
    FOREIGN KEY (location_id) REFERENCES dim_location (location_id),
    FOREIGN KEY (company_id) REFERENCES dim_company (company_id)
);
```

Tabla de hechos

```sql
CREATE TABLE dim_company (
    company_id SERIAL PRIMARY KEY,
    company VARCHAR(255),
    industry VARCHAR(100),
    stage VARCHAR(100)
);
```

Dimensión de datos de Compañía

```sql
CREATE TABLE dim_location (
    location_id SERIAL PRIMARY KEY,
    location_hq VARCHAR(255),
    country VARCHAR(100),
    continent VARCHAR(100),
    latitude DECIMAL(9,6),
    longitude DECIMAL(9,6)
);
```

Dimensión de datos geográficos

```sql
CREATE TABLE dim_time (
    date_id INT PRIMARY KEY,
    date DATE,
    year INT
);
```

Dimensión de datos temporales

Ejemplos queries modelo dimensional:

Trends de despidos por año:

```
SELECT dim_time.year, COUNT(fact_layoffs.fact_id) AS
number_of_layoff_events, SUM(fact_layoffs.laid_off) AS total_laid_off
FROM fact_layoffs
JOIN dim_time ON fact_layoffs.date_id = dim_time.date_id
GROUP BY dim_time.year
ORDER BY dim_time.year;
```

Resultado:

year	number_of_layoff_events	total_laid_off
2020	294	57724
2021	10	3120
2022	475	110830
2023	377	138219
2024	12	2150

Número total de despidos por industria:

```
SELECT dim_company.industry, SUM(fact_layoffs.laid_off) AS
total_laid_off
FROM fact_layoffs
JOIN dim_company ON fact_layoffs.company_id = dim_company.company_id
GROUP BY dim_company.industry;
```

Resultado:

industry	total_laid_off
Real Estate	9954
Manufacturing	790
Healthcare	12013
Construction	1089
Media	6002
Consumer	48733

Trends de despidos por región:

```
SELECT dim_location.continent, dim_location.country,
SUM(fact_layoffs.laid_off) AS total_laid_off
FROM fact_layoffs
JOIN dim_location ON fact_layoffs.location_id = dim_location.location_id
GROUP BY dim_location.continent, dim_location.country
ORDER BY dim_location.continent, dim_location.country;
```

Resultado:

continent	country	total_laid_off
Africa	Kenya	982
Africa	Nigeria	950
Africa	Senegal	300
Asia	China	180
Asia	Hong Kong	700
Asia	India	21836

Conclusiones (Dashboard y trabajo realizado del Modelo dimensional)

- La industria con más registros de despidos en Estados Unidos es la industria de las finanzas (15%). Con una diferencia de registros del 6% en comparación con el segundo puesto obtenido, la industria minorista (8,6%).
- La industria con más número total de despidos en Estados Unidos es la industria del consumidor con más de 4.000 despidos, esta industria hace referencia al sector que produce bienes y servicios destinados al consumo por parte de los individuos (servicios, bienes duraderos y bienes no duraderos).
- La industria minorista se consolida como la segunda industria con más número total de despidos y más número de registro de despidos, sugiriendo que esta industria ha estado sufriendo muchos cambios o desafíos a lo largo del 2020 a 2023. Esto por diferentes factores que no podemos dar como ciertos, pero recordemos que estamos abarcado años donde la pandemia seguía presente en este país y industrias como estas tuvieron que adaptarse a nuevas formas de ventas que no fueran físicas.
- El año 2023 lidera como el año con más números de despidos casi en todas las industrias de Estados Unidos, seguido por el año 2022 que está presente con más despidos en industrias de comida, salud y hardware. Lo curioso de esto es que el año 2022 es el que lidera con más número de registros en Estados Unidos.
- Podemos ver como el tamaño de las industrias antes y después de los despidos, no cambia drásticamente, manteniendo la mayoría de industrias su mismo “puesto”, pero tenemos casos como la industria de las ventas la cual su tamaño termina siendo inferior al de la industria de transporte después de los despidos que esta sufrió.

Conclusiones (Dashboard API y Realización del EDA)

- La distribución de empresas por país en el mercado se concentra principalmente en cuatro. China con un 23.2%, Estados Unidos con un 19,9%, India con un 19,8%, y Japón con un 16,4%.
- Después de esos países la distribución entre países se vuelve muy pequeña, comenzando con el quinto puesto que se lo lleva reino unido con tan solo un 4.7% seguido por hong kong con un 4.2%
- Podemos ver cómo a diferencia de en nuestro dataset de despidos en el que más de un 60% de los registros eran de estados unidos, en este dataset de la API podemos ver una tendencia no tan abrupta como con nuestro dataset, teniendo la API una distribución entre varios países y no una tendencia directa en solo un país. Pero claro que entre estos cuatro países sigue apareciendo nuestro país predominante en nuestro dataset de despidos, Estados unidos.
- Si hablamos de la distribución de como es la industria en nuestro dataset de mercado, encontramos que la industria que predomina es la de bancos regionales con un 18.9% seguida por industrias especializadas e ingenierías como la maquinaria o la química
- Las industrias con menos presencia en el mercado son las de seguros con tan solo 20-25 registros seguidas por varias industrias que hacen referencia a subcategorías de lo que es la industria de bienes raíces, en total tenemos 4 subcategorías, inversiones de bienes raíces en oficinas, hoteles, centros de salud y de nicho. Todas estas “industrias” tienen alrededor de solamente 20-13 compañías registradas en nuestro dataset de la API.
- Si hablamos de la industria con mayor margen de beneficio según nuestro dataset de mercado, esta es la industria de entretenimiento siendo realmente más rentable que todas las demás industrias del dataset, sorprendiéndonos con una margen de beneficio del 57% en comparación con el margen de beneficio de las demás industrias.
- Los países con más número de empleados a tiempo real son Estados Unidos con un 31.5%, China con un 18.4%, Japón con un 11.8% e India con un 8.6%.
- Aunque vemos los mismos países que predominan en el mercado, aquí intercambian papeles y es que Estados Unidos termina siendo el país con más número de empleados a tiempo real, seguido por China.
- Si analizamos un panorama general de la distribución de empleados por industria podemos ver como industrias como la de bancos, salud y servicios tecnológicos son los que predominan, esto es lógico y es que, si pensamos cómo es que los servicios que ofrecen esta industria son los más necesarios para la vida dentro de una sociedad y mercado, es normal que su número de empleados a tiempo real sea muy alto, ya que son servicios básicos que se necesitan la mayoría las 24 horas del dia.
- Es interesante analizar que en el panorama general de la distribución de empleados por industria, la industria de bancos internacionales y regionales salgan a la luz, dándonos más luz a poder sacar conclusiones como la que dijimos en el anterior punto.