Entrega Final - Proyecto ETL
Datadive - Tech Layoffs

Realizado por:
Federico Gonzalez Cardenas
Isaac Piedrahita Carvajal

Docente
Javier Alejandro Vergara Zorrilla
ETL

Programa de Ingeniería de Datos e Inteligencia Artificial
Facultad de ingeniería
Cuarto Semestre
Universidad Autónoma de Occidente
Santiago de Cali
2024

Aspectos a tener en cuenta:

El video evidencia, llamado **Real Time Dashboard ETL** está enfocado en mostrar el proceso de kafka stream a un dashboard en tiempo real, específicamente en este video se puede visualizar que este streaming se está corriendo en otra arquitectura que no es la del proyecto,está diseñada de esa manera por temas de tiempo y para la presentación, de igual manera la colocamos en el documento ya que mostrar el comportamiento real de las gráficas con las que trabajamos nos parece fundamental y enriquecedor

El video **Demostracion-Airflow** Muestra todo el proceso de Airflow solicitado, hasta el kafka producer, se muestra después la llegada a los datos al kafka consumer y el envío de los datos a tiempo real a nuestro reporte en Power BI, Un reporte con unas graficas de muestra, que **por eso quisimos dejar como evidencia el video de Real Time Dashboard ETL**, sin dejar de lado que **el video de Demostracion-Airflow se ejecuta en la arquitectura de carpetas del proyecto**, cosa que el otro video no hace.

La carpeta de **Great_Expectations** contiene los archivos de las expectativas hechas para el dataset de la APi y nuestro dataset original de despidos que elegimos desde el primer corte, estas expectativas se hicieron ya limpiados los datasets, en busca de crear expectativas que logren indicarnos que nuestra limpieza fue exitosa segun los criterios con los cuales decidimos hacer esta limpieza, Cada dataset tiene su propio script, y cada script tiene su propio reporte json que busca mostrar todas las expectativas que se le hicieron a cada dataset, de igual manera en este documento mostraremos como al codigo le añadimos otro reporte mas especifico.

Proceso merge:

Desafíos:
Falta de Correspondencia Directa entre Compañías: Intentamos hacer un match directo entre las empresas del dataset de despidos y las empresas del dataset de la API, pero descubrimos que estos datasets no compartían compañías en común.
Diferencias en la Nomenclatura de Industrias: Las industrias en ambos datasets estaban etiquetadas de manera diferente, lo que impedía una correspondencia directa y simple.

Solución Implementada:
Dado que no había coincidencias directas entre las compañías en los dos datasets, optamos por enriquecer el dataset de despidos mediante el mapeo de las industrias. Calculamos las métricas financieras promedio por industria a partir del dataset de la API y asignamos estas métricas a las industrias mapeadas en el dataset de despidos.

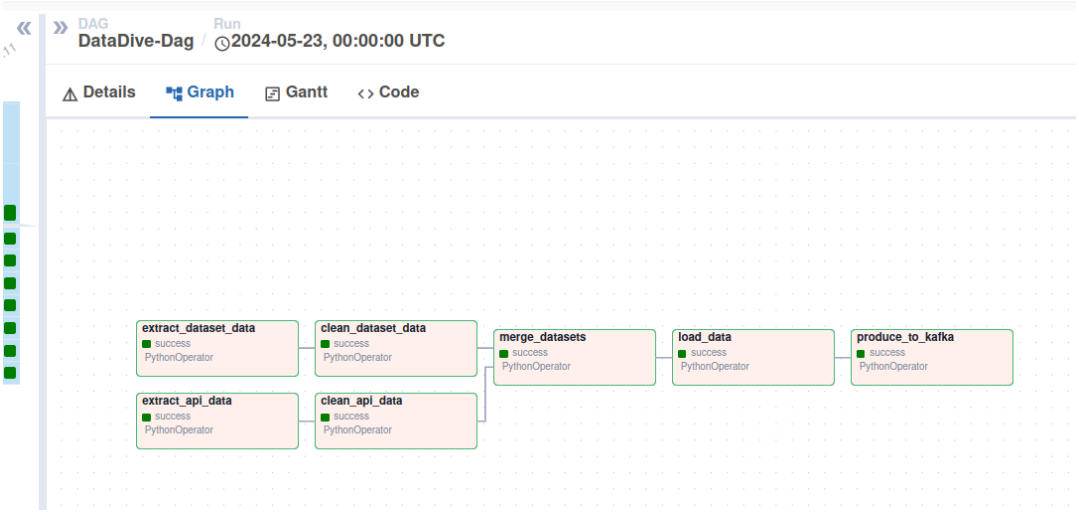
Este proceso nos permitió proporcionar una visión más completa y detallada de las industrias afectadas por los despidos. Además un análisis más profundo, otra manera de mirar las razones de los despidos y una toma de decisiones más informada basada en la combinación de datos financieros e información de despidos.

Vídeos evidencia:

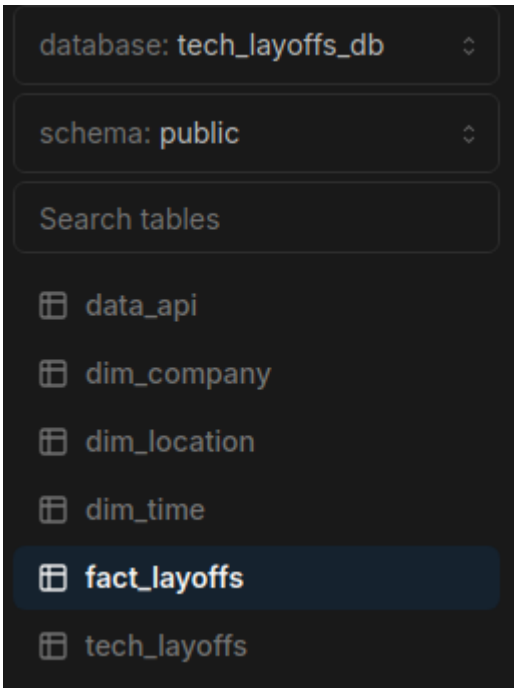
Real Time Dashboard ETL - Tech Layoffs: https://www.youtube.com/watch?v=9JB5zr0uge8

Demostracion-Airflow: https://www.youtube.com/watch?v=JDDz427RQGI

Evidencias gráficas:
Airflow Success

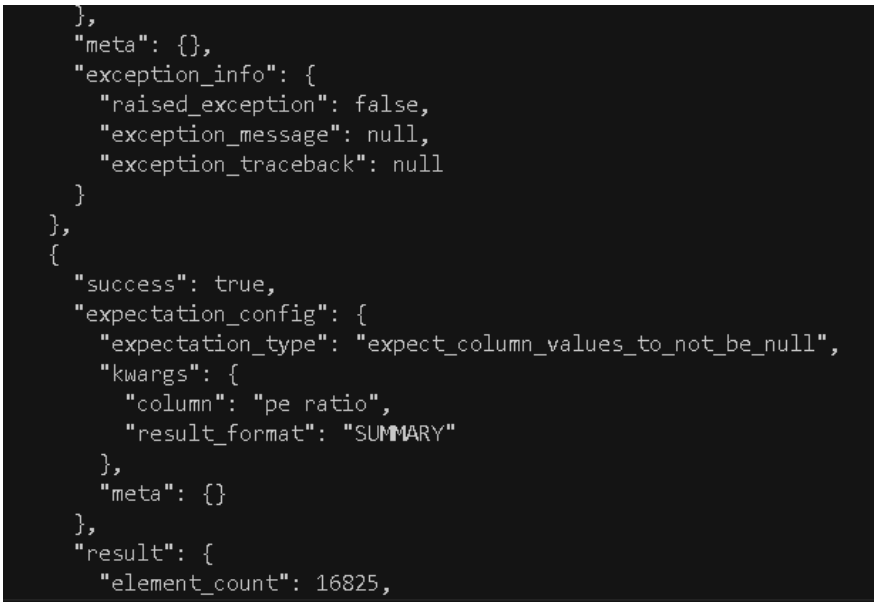


Tablas DB:



Evidencia Great Expectations (Se quiere mostrar parte del informe que sacamos detallado de cada dataset):

Parte del informe detallado (hay dos informes, uno el cual estará en un archivo json en el github y otro mucho más detallado que es el que añadimos nosotros al código, pero este solo lo imprime en la terminal, ya que es un informe muy extenso) de algunas expectativas trabajadas para los datos de la API ya transformados, en total se hicieron 14 para la API queriéndonos enfocar en expectativas tanto generales como específicos para el comportamiento esperado en campos específicos




```
    ]
  },
  "meta": {},
  "exception_info": {
    "raised_exception": false,
    "exception_message": null,
    "exception_traceback": null
  }
},
{
  "success": true,
  "expectation_config": {
    "expectation_type": "expect_column_values_to_be_of_type",
    "kwargs": {
      "column": "pe ratio",
      "type_": "float",
      "result_format": "SUMMARY"
    },
    "meta": {}
  },
  "result": {
    "observed_value": "float64"
  },
  "meta": {},
  "exception_info": {
    "raised_exception": false,
    "exception_message": null,
    "exception_traceback": null
  }
},
{
  "success": true,
  "expectation_config": {
    "expectation_type": "expect_column_values_to_be_between",
    "kwargs": {
      "column": "pe ratio",
      "min_value": 0,
      "max_value": null,
      "result_format": "SUMMARY"
    },
```



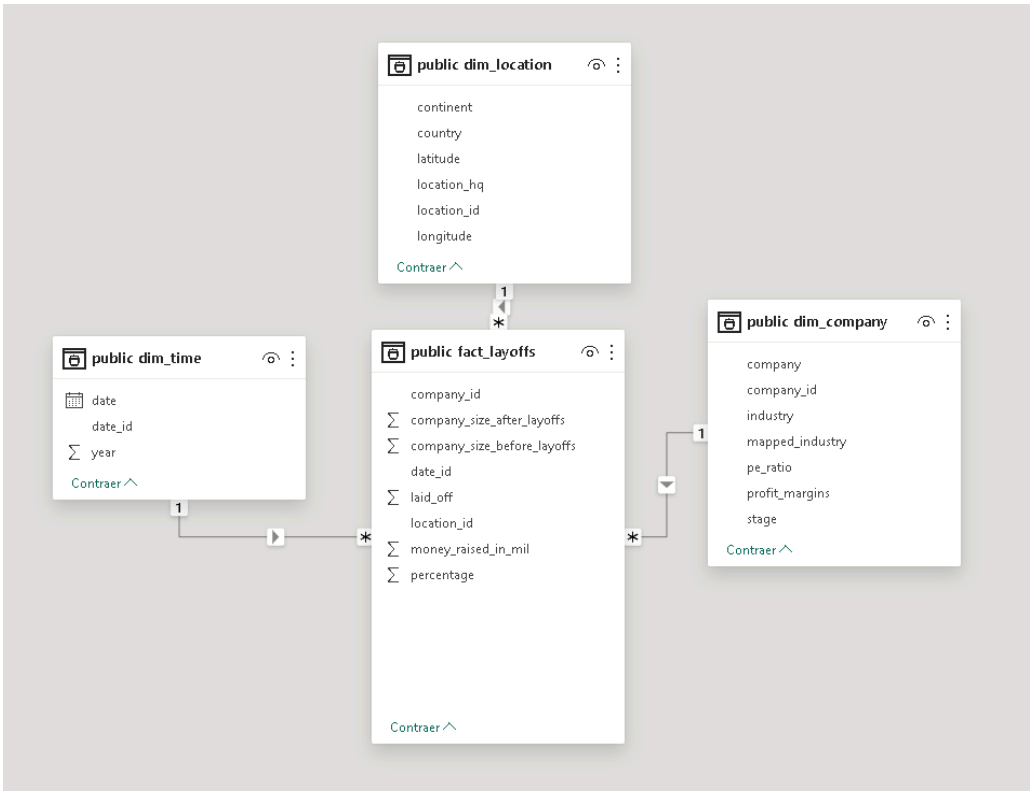
```
{
  "success": true,
  "expectation_config": {
    "expectation_type": "expect_column_values_to_be_between",
    "kwargs": {
      "column": "full time employees",
      "min_value": 0,
      "max_value": null,
      "result_format": "SUMMARY"
    },
    "meta": {}
  },
  "result": {
    "element_count": 16825,
    "missing_count": 0,
    "missing_percent": 0.0,
    "unexpected_count": 0,
    "unexpected_percent": 0.0,
    "unexpected_percent_total": 0.0,
    "unexpected_percent_nonmissing": 0.0,
    "partial_unexpected_list": [],
    "partial_unexpected_index_list": [],
    "partial_unexpected_counts": []
  },
```


Para la tabla de TechLayoffs se hicieron un total de 5 expectativas buscando mantener expectativas generales y específicas enfocadas en rangos o conjuntos específicos que abarquen los datos

```
    },
    {
      "success": true,
      "expectation_config": {
        "expectation_type": "expect_column_mean_to_be_between",
        "kwargs": {
          "column": "money_raised_in_mil",
          "min_value": 0,
          "max_value": 894.09,
          "result_format": "SUMMARY"
        },
        "meta": {}
      },
      "result": {
        "observed_value": 894.0873287671233,
        "element_count": 1168,
        "missing_count": null,
        "missing_percent": null
      },
      "meta": {},
      "exception_info": {
        "raised_exception": false,
        "exception_message": null,
        "exception_traceback": null
      }
    }
  ],
  "evaluation_parameters": {},
  "statistics": {
    "evaluated_expectations": 5,
    "successful_expectations": 4,
    "unsuccessful_expectations": 1,
    "success_percent": 80.0
  },
  "meta": {
    "great_expectations_version": "0.18.13",
    "expectation_suite_name": "default",
    "run_id": {
```

```
{
  "success": true,
  "expectation_config": {
    "expectation_type": "expect_column_values_to_not_be_null",
    "kwargs": {
      "column": "company",
      "result_format": "SUMMARY"
    },
    "meta": {}
  },
  "result": {
    "element_count": 1168,
    "unexpected_count": 0,
    "unexpected_percent": 0.0,
    "unexpected_percent_total": 0.0,
    "partial_unexpected_list": []
  },
  "meta": {},
  "exception_info": {
    "raised_exception": false,
    "exception_message": null,
    "exception_traceback": null
  }
},
{
  "success": true,
  "expectation_config": {
    "expectation_type": "expect_column_values_to_not_be_null",
    "kwargs": {
      "column": "country",
      "result_format": "SUMMARY"
    },
    "meta": {}
  },
  "result": {
    "element_count": 1168,
```

Nuevo modelo dimensional (Campos del dataset de la API añadidos)



Conclusiones dashboard en tiempo real:

-El enfoque tomado en este dashboard es mostrar la diferencia que existe en nuestro dataset entre los registros de despidos y los despidos totales de cada industria. Demostrando que en la mayoría de casos las industrias con más registros de despidos, sean las industrias con mayores despidos totales, aunque pueda parecer a un primer análisis que tienen una relación, es correcto ya que las industrias que veremos con mayor número de registros, también en su mayoría las podremos ver que tienen un valor alto de despidos totales, pero que no necesariamente haya una relación 1 a 1 y que por que la industria con mayor número de registros de despidos sea la industria de finanzas, tenga que cumplir con que también esa misma industria sea la que tenga un mayor número de despidos totales, y esta es la conclusión que queremos demostrar con este dashboard principalmente.

-Trabajamos también con valores referentes al tamaño de las industrias antes y después de los despidos, queriendo hacer notar que en su gran mayoría no hay un gran cambio y que las industrias se mantienen en su mismo posicionamiento de tamaño **a diferencia de otros** tanto antes como después de los despidos, aunque claro que tenemos casos en los que el tamaño de la industria baja lo suficiente como para que otra tome su lugar después de los despidos y este caso es el de la industria de **Transportation** el cual tenía un tamaño superior a la industria de **Sales** antes de los despidos, cosa que después de los despidos estas industrias intercambian papeles perjudicando lo suficiente a la industria de **Transportation** como para que termine teniendo un tamaño inferior a la industria de **Sales**.

-En la última página de nuestro dashboard quisimos mostrar los dos valores que marcan la diferencia de la cual hablamos, mostrando el **número total de registros** y el **número de despidos totales a tiempo real**, ya que mostrando números referentes a estos dos valores hace que se cumplan y se vean dos requisitos los cuales son el de mostrar visualmente que se están enviando los datos de nuestro merge y mostrando con valores numéricos el enfoque de nuestro dashboard.

Conclusiones dashboard Merge:
Aquí tienes algunas conclusiones concisas del dashboard adjunto:

Gráfica #1: Promedio de Dinero Recaudado por Industria (en millones):
- La industria de consumo tiene el mayor promedio de dinero recaudado, con 972.27 millones.
- Las industrias de bienes raíces y salud también recaudan cantidades significativas, con 775.53 millones y 341.11 millones respectivamente.
- La industria de recursos humanos tiene el promedio más bajo con 177.50 millones.

Gráfica #2: Promedio de PE Ratio por Industria:
- La industria de salud tiene el PE Ratio promedio más alto, con 286.19. Esto indica que las acciones de esta industria están muy peleadas entre los inversionistas, y se espera que estas tengan avances considerables en el porvenir.
- Las industrias de retail y bienes raíces también tienen PE Ratios altos, con 193.59 y 187.52 respectivamente.
- La industria de energía tiene el PE Ratio más bajo, con 104.48. Por lo que los inversores están dispuestos a pagar en promedio 104.48$ Por cada dólar de ganancia respecto al precio de la acción, esto nos indica que las acciones de esta industria no se encuentran peleadas en el momento.

Gráfica #3: Promedio de PE Ratio y Márgenes de Ganancia por Industria:
- Hay una alta variabilidad en el PE Ratio, especialmente en la industria de salud.
- Los márgenes de ganancia son más altos en la industria de energía, a pesar de tener un PE Ratio más bajo.

Gráfico #4: Relación entre PE Ratio y Márgenes de Ganancia por Industria:
- Observamos que hay una relación directa y clara entre el PE Ratio y los márgenes de ganancia, pero se observa que algunas industrias con PE Ratios más bajos, como la energía, tienen márgenes de ganancia más altos, esto nos podría indicar que Industrias con PE Ratio alto pero Márgenes de ganancia bajos podrían encontrarse sobrevaloradas en el mercado.

¿Qué significa PE Ratio y Profit Margin?

- **PE Ratio (Price-to-Earnings Ratio)**: Es una medida de valoración que compara el precio de las acciones de una empresa con sus ganancias por acción (EPS). Un PE Ratio alto puede indicar que las acciones están **sobrevaloradas** o que los **inversores esperan un alto crecimiento futuro.**

- **Profit Margin (Margen de Ganancia)**: Nos indica la rentabilidad que muestra el porcentaje de ingresos que una empresa conserva como **ganancias** después de todos los costos y gastos. Un margen de ganancia más alto nos indica una mayor eficiencia en la gestión de costos y mayor rentabilidad.