REPORTE DE RESULTADOS

Ing. Hilaría Adima Vásquez Durán

Sede La Paz - Bolivia

OBJETIVOS DEL ANALISIS DE DATOS

- Se debe realizar la migración masiva de datos de MySQL a **HIVE HADOOP** para que deje de presentar fallas debido a que MySQL su diseño y el aumento en el volumen de datos.
- Elaborar el análisis de datos para detectar oportunidades de negocio en la dinámica de aviación en Estados Unidos.
- Utilizar el proceso ETL(Extract, Transform, Load) para la migración de información del primer trimestre del año 2024.
- Desarrollar scripts en Python y en HiveQL que demuestren los 10 mejores, promedio de vuelos, aeropuertos de los que parte más vuelos, las rutas de vuelo que sufren mayores demoras.

ARQUITECTURA DE LA MIGRACIÓN HACIA CÓMPUTO DISTRIBUIDO

- El análisis utilizado para la migración de datos se utilizó el proceso ETL (Extract, Transform, Load) .Se realizó lo siguiente:

Se realizo la extracción de la información de la base de datos, que se encuentra diseñada en MySQL, del sistema de origen.

Una vez que se obtiene la información, se realiza una verificación sobre los datos obtenidos si se encuentran correctos .

Los procesos ETL heredados importan los datos, los limpian la información que no localmente y, a continuación, los almacenan en un motor de datos relacionales.

ANÁLISIS E INTERPRETACIÓN DE DATOS

- Se realizo el análisis de la migración de datos para poder obtener los requerimientos de la empresa de Aero Líneas, llegando a los siguientes resultados:

El análisis utilizado para la migración de datos se utilizó el proceso ETL (Extract, Transform, Load)
Se realizó lo siguiente:

Extraer los datos desde los sistemas de origen.

Analizar los datos extraídos obteniendo con verificación.

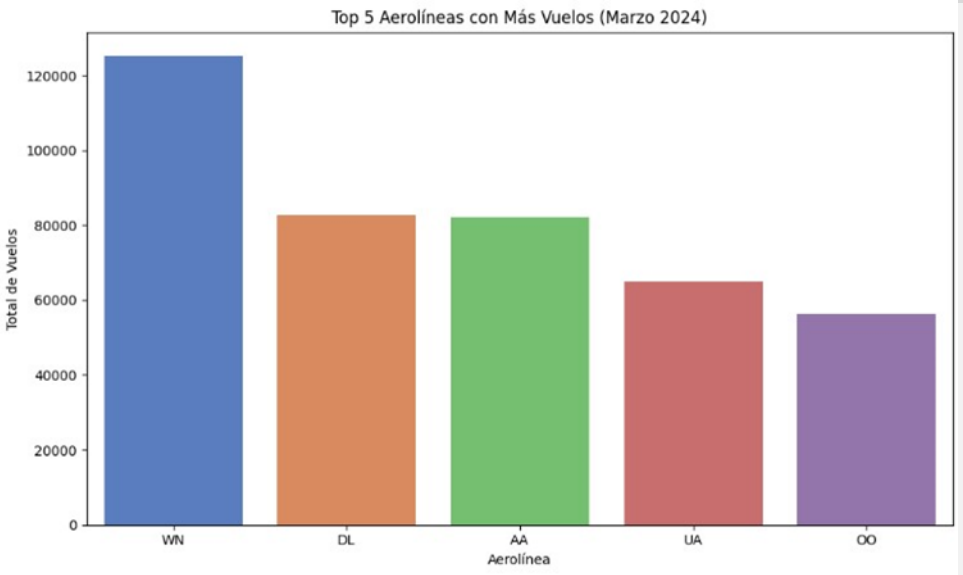
Convertir los datos de MySQL en base de datos distribuidas para la transformación se utilizo el FRAMEWORK HIVE HADOOP

ANÁLISIS E INTERPRETACIÓN DE DATOS

Realizando La evaluación sobre el valor máximo de las 5 aerolíneas con mas vuelos son:

```
Top 5 aerolíneas con más vuelos:
   Reporting_Airline  Total_Flights
13                WN         125272
4                 DL          82668
1                 AA          82259
12                UA          64929
11                OO          56241
<ipython-input-55-9461a613a2a7>:26: FutureWarning:
```

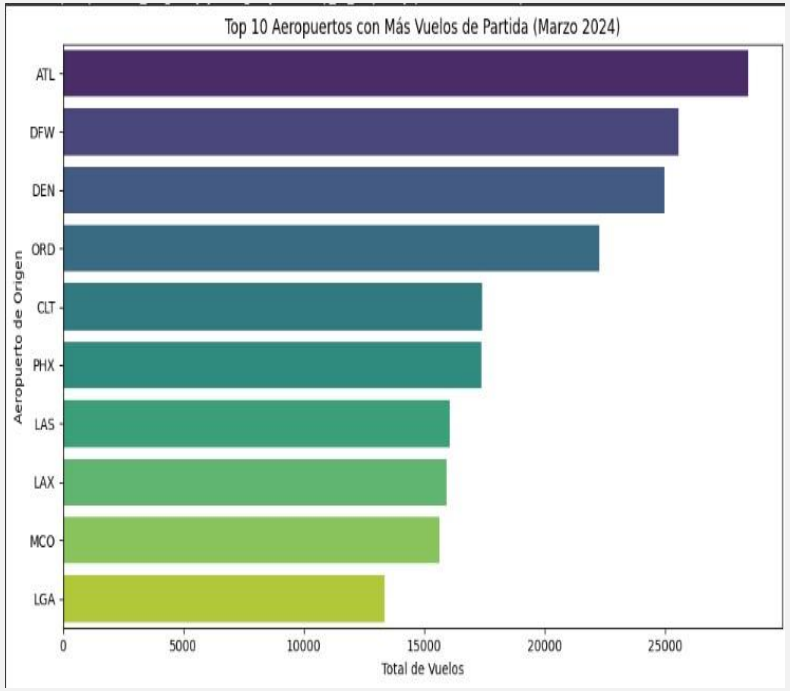
Como podemos observar en la siguiente gráfica que la aerolínea con más vuelos es Northwest Airlines Inc.



ANÁLISIS E INTERPRETACIÓN DE DATOS

Como se realizó de las aerolíneas, también se efectuó de los aeropuertos de los que parte más vuelos

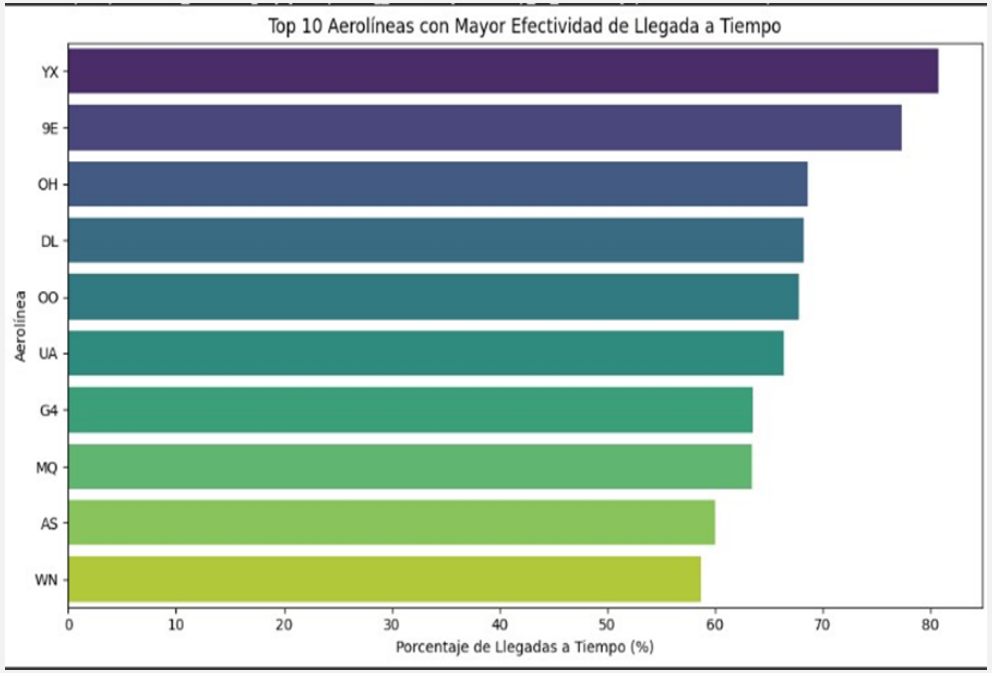
```
Top 10 aeropuertos con más vuelos de partida:
    Origin  Total_Flights
18     ATL          28459
87     DFW          25567
86     DEN          24976
225    ORD          22288
67     CLT          17400
236    PHX          17375
171    LAS          16062
173    LAX          15916
192    MCO          15650
182    LGA          13366
<ipython-input-63-b12285c73510>:24: FutureWarning:
```



ANÁLISIS E INTERPRETACIÓN DE DATOS

También se evaluó las afluencia de las líneas Aéreas que el cliente utiliza donde se expresa cuales son los más altas su servicio.

```
Top 10 aerolíneas con mayor efectividad de llegada a tiempo:
   Reporting_Airline  OnTime_Percentage
14                YX          80.782816
0                 9E          77.318017
10                OH          68.624741
4                 DL          68.280350
11                OO          67.811739
12                UA          66.361718
6                 G4          63.520986
8                 MQ          63.463146
2                 AS          59.992871
13                WN          58.747366
<ipython-input-57-dc05966f0ada>:32: FutureWarning:
```



ANÁLISIS E INTERPRETACIÓN DE DATOS

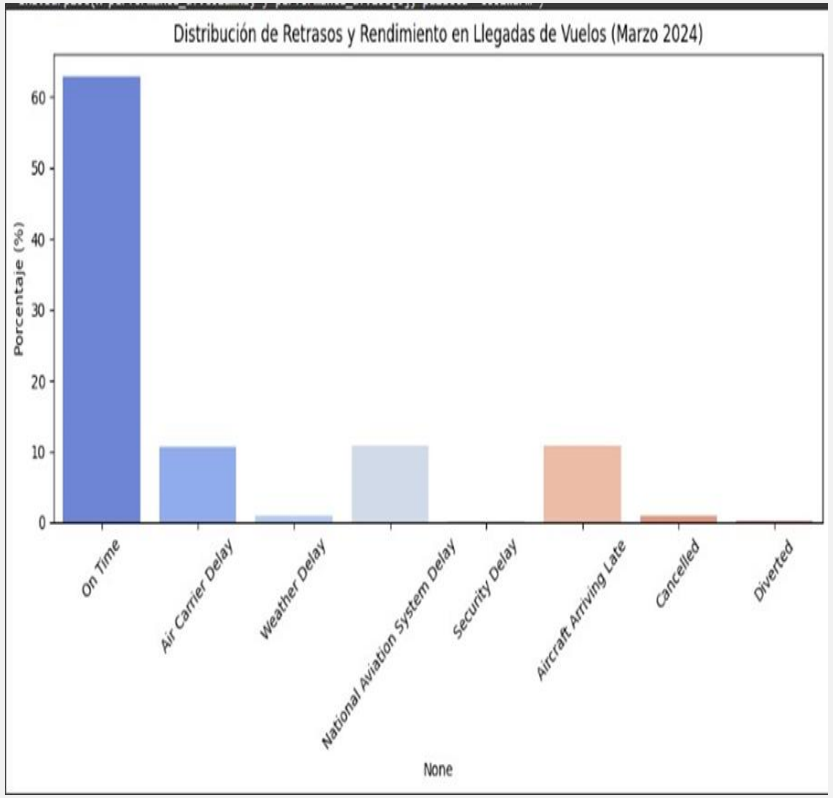
Se realizó la distribución del performance de llegada de los vuelos.

```
   Year  Quarter  Month  DayofMonth  DayOfWeek  FlightDate Reporting_Airline  \
0  2024        1      3           1          5  2024-03-01                9E
1  2024        1      3           2          6  2024-03-02                9E
2  2024        1      3           3          7  2024-03-03                9E
3  2024        1      3           4          1  2024-03-04                9E
4  2024        1      3           5          2  2024-03-05                9E

   DOT_ID_Reporting_Airline IATA_CODE_Reporting_Airline Tail_Number  ...  \
0                     20363                         9E      N935XJ  ...
1                     20363                         9E      N910XJ  ...
2                     20363                         9E      N298PQ  ...
3                     20363                         9E      N602LR  ...
4                     20363                         9E      N348PQ  ...

   Div4TailNum  Div5Airport  Div5AirportID  Div5AirportSeqID Div5WheelsOn  \
0          NaN          NaN            NaN               NaN          NaN
1          NaN          NaN            NaN               NaN          NaN
2          NaN          NaN            NaN               NaN          NaN
3          NaN          NaN            NaN               NaN          NaN
4          NaN          NaN            NaN               NaN          NaN

   Div5TotalGTime Div5LongestGTime  Div5WheelsOff Div5TailNum  Unnamed: 109
0             NaN              NaN            NaN         NaN           NaN
1             NaN              NaN            NaN         NaN           NaN
2             NaN              NaN            NaN         NaN           NaN
3             NaN              NaN            NaN         NaN           NaN
4             NaN              NaN            NaN         NaN           NaN
```



CONCLUSIONES

Al realizar la categorización de datos se logró agrupar los datos en una base distribuida para poder mostrar los vuelos que se realizan de acuerdo a cada aerolínea y no se tengan problemas en el momento de acceder a la información. Se logró la migración de datos en una base distribuida para poder realizar el análisis de los vuelos aerolíneas y aeropuertos no se tengan problemas en el momento de acceder a la información.

La utilización de la migración del framework Hadoop sirvió para poder generar DATAMARKS de la información DEL DATA WERHOUSE de acuerdo a la cantidad de vuelos en la clasificación trimestral.

Con el análisis realizado se demostró que la información es obtenida en tiempo real de acuerdo a la necesidad que tiene el usuario.

Categorización de cuerdo a Aeropuertos más de alto tráfico en los Estados Unidos

LIMITACIONES

- Los obstáculos a presentarse, que MySQL es un gestor de base de datos cuanto mayor información se tiene almacenada es mucho más lento.

- Contar con el personal o software que realice el análisis trimestral de los vuelos para la toma de decisiones de aquellas aerolíneas que tienen más frecuencia en en volar versus las que no vuelan seguido.

Gracias

Hilaría Adima Vásquez Duran

591 65533537

havasquezd710316@Gmail,com

GitHub - HAVD-2024/Manejo-eficiente-de-datos-al-programar-aplicaciones-distribuidas