

BIG DATA

PROYECTO FINAL

Atanas Turlakov



Atanas Turlakov

RESPONSABLE DEL DESARROLLO DE LA APLICACIÓN

Fecha: 11.02.2024



Resumen:

El objetivo de este estudio es realizar un analisis del aeropuerto de San Francisco y en particular el trafico saliente hacia Tokio y otras ciudades de Japon. Se va analizar un archivo con datos de todo el trafico hacia estos destinos realizando varias tareas que ayudaran a sacar conclusiones y tomar deseciones a base de ellas.

Metodologia:

En este estudio se van a usar varias herramientas de programacion y analisis de datos que ayudaran a sacar conclusiones mas precisas

- Python: Language de programacion de alto nivel y multiplataforma
- PySpark: Biblioteca de Python para trabajar con Apache Spark
- Bash scripting: Programacion en entornos Linux/UNIX. Con esta herramienta se va a realizar le preparacion de alguno de los entornos de ejecucion
- Docker: Plataforma para crear y ejecutar aplicaciones aisladas del host
- Apache Cassandra: Base de datos NoSQL
- CSV: Formato de archivo para organizar datos de manera sencilla
- Debian 12 GNU/Linux – Sistema operativo en que estan probados los scripts



Desarrollo y resultados obtenidos:

En el primer estudio que realizaremos es una categorización de los datos por su tipo lo que nos permitira tener una imagen amplia para seguir adelante con el analisis. Obtenemos una tabla con el siguiente resultado:

Nombre del Campo	Tipo de Dato
Activity Period	Numérico
Operating Airline	Categórico
Operating Airline IATA Code	Categórico
Published Airline	Categórico
Published Airline IATA Code	Categórico
GEO Summary	Categórico
GEO Region	Categórico
Activity Type Code	Categórico
Price Category Code	Categórico
Terminal	Categórico
Boarding Area	Categórico
Passenger Count	Numérico
Adjusted Activity Type Code	Categórico
Adjusted Passenger Count	Numérico
Year	Numérico
Month	Categórico

Aqui nos muestra que tenemos 4 campos numericos y el resto son categoricos.

En el siguiente estudio se ejecutan consultas sobre los registros de Air China y Air Berlin.

Recuperar todos los registros de “Air China”:

Recuperar todos los registros de la aerolínea “Air China”

Published Airline	Published Airline IATA Code	Terminal	Boarding Area	Passenger Count	Month	Year	GEO Region	Price Category Code	Activity Type Code
Air China	CA	International	G	7187	March	2011	Asia	Other	Deplaned
Air China	CA	International	G	7154	May	2007	Asia	Other	Deplaned
Air China	CA	International	G	2767	February	2006	Asia	Other	Enplaned
Air China	CA	International	G	8187	September	2014	Asia	Other	Deplaned
Air China	CA	International	G	7467	December	2012	Asia	Other	Enplaned
Air China	CA	International	G	7836	March	2015	Asia	Other	Enplaned
Air China	CA	International	G	5596	November	2008	Asia	Other	Deplaned
Air China	CA	International	G	7659	March	2014	Asia	Other	Deplaned
Air China	CA	International	G	7188	November	2011	Asia	Other	Deplaned
Air China	CA	International	G	6693	November	2011	Asia	Other	Enplaned
Air China	CA	International	G	7653	August	2012	Asia	Other	Enplaned
Air China	CA	International	G	6883	March	2012	Asia	Other	Enplaned
Air China	CA	International	G	3078	January	2006	Asia	Other	Enplaned
Air China	CA	International	G	7484	December	2011	Asia	Other	Enplaned
Air China	CA	International	G	7078	April	2012	Asia	Other	Deplaned
Air China	CA	International	G	2862	February	2006	Asia	Other	Deplaned
Air China	CA	International	G	7446	October	2014	Asia	Other	Enplaned
Air China	CA	International	G	8771	July	2013	Asia	Other	Deplaned
Air China	CA	International	G	3800	February	2009	Asia	Other	Enplaned

Aquí se pueden sacar conclusiones sobre las temporadas. Se ve claramente que la temporada baja se sentra sobre todo en los meses de enero y febrero. Basado en ello la compania o el aeropuerto puede ajustar el numero de vuelos y ventas basado a esta informacion

Recuperar todos los registros de "Air Berlin" embarcados por la puerta "G":

```
##### Recuperar todos los vuelos de la compañía "AirBerlin"embarcados porla puerta "G" #####
```

Published Airline	Published Airline IATA Code	Terminal	Boarding Area	Passenger Count	Month	Year	GEO Region	Price Category Code	Activity Type Code
Air Berlin	AB	International	G	2357	September	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	2620	July	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	2085	August	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	2343	September	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	1455	October	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	2261	July	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	972	May	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	2294	August	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	1337	May	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	2540	June	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	1689	October	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	2150	June	2010	Europe	Other	Deplaned

```
hasko@desktop:~$
```

En el resultado se puede observar que que Air Berlin ha tenido bajo volumen de pasajeros embarcados por la puerta G. Tambien que esta ha sido solo en año 2010 que se puede tratar de un caso aislado como situaciones puntuales que han obligado a redirigir trafico hacia esta puerta


Las siguientes tareas se van a realizar con Apache Spark usando Python con la librería pyspark editados con cuadernos Google Colab

Lo primero sacaremos el numero de todas las companias operando en el aeropuerto de San Francisco:

```
+-----+  
| Operating Airline|  
+-----+  
|      Icelandair|  
|      Ameriflight|  
|      Cathay Pacific|  
|      Aeromexico|  
|      Etihad Airways|  
|      Philippine Airlines|  
|      United Airlines -...|  
|      Turkish Airlines|  
|      Swiss International|  
|      Independence Air|  
|      Miami Air Interna...|  
|      Air France|  
|      Japan Airlines|  
|      Midwest Airlines|  
|      Atlas Air, Inc|  
|      JetBlue Airways |  
|      China Eastern|  
|      Mexicana Airlines|  
|      Air Canada |  
|      Allegiant Air|  
+-----+  
only showing top 20 rows  
  
En el registro hay 77 diferentes companias
```

Se puede ver claramente que en el aeropuerto operan 77 companias diferentes

La siguiente consulta es sobre la media de pasajeros de cada compania




```
+-----+-----+
| Operating Airline|avg(Passenger Count)|
+-----+-----+
|      Icelandair|      2,799.70|
|      Ameriflight|         5.00|
|      Cathay Pacific|    17,121.33|
|      Aeromexico|     5,463.82|
|      Etihad Airways|    6,476.09|
|      Philippine Airlines|  10,248.64|
|      United Airlines -...|  48,915.47|
|      Turkish Airlines|     8,162.42|
|      Swiss International|    6,061.64|
|      Independence Air|    6,391.30|
|      Miami Air Interna...|     107.38|
|      Air France|    11,589.08|
|      Japan Airlines|     6,470.33|
|      Midwest Airlines|     3,883.00|
|      Atlas Air, Inc|         34.00|
|      JetBlue Airways |    35,261.14|
|      China Eastern|     5,498.40|
|      Mexicana Airlines|    7,993.81|
|      Air Canada |    18,251.56|
|      Allegiant Air|     1,516.81|
+-----+-----+
only showing top 20 rows
```

De esta consulta se pueden sacar conclusiones sobre la carga de cada compania y tomar deseciones futuras sobre tasas de aeropuerto, tasas de carga de equipaje etc.



Eliminaremos los registros duplicados por el campo "GEO Región", manteniendo

únicamente aquel con mayor número de pasajeros.



GEO Region	Max Passenger Count
Europe	48136
Central America	8970
US	659837
South America	3685
Mexico	29206
Middle East	14769
Canada	39798
Australia / Oceania	12973
Asia	86398

En esta consulta se puede observar el numero de pasajeros saliendo a cada region del mundo. Se ve que la region de Estados Unidos sigue teniendo el mayor numero debido a la distancia y numero de vuelos y la region de Asia en segundo puesto debido a la poblacion del continente ya que tiene 60% de la poblacion mundial

La siguiente tarea se usara PySpark para hacer un analisis descriptivo para calcular la media de desviacion estandar de cada elemento del conjunto de datos numericos

```
➡ +-----+-----+
|mean_Activity Period|stddev_Activity Period|
+-----+-----+
| 201045.07336576266| 313.33619609986414|
+-----+-----+

+-----+-----+
|mean_Passenger Count|stddev_Passenger Count|
+-----+-----+
| 29240.521090157927| 58319.509284123524|
+-----+-----+

+-----+-----+
|mean_Adjusted Passenger Count|stddev_Adjusted Passenger Count|
+-----+-----+
| 29331.917105350836| 58284.1822186625|
+-----+-----+

+-----+-----+
| mean_year| stddev_year|
+-----+-----+
|2010.385220230559|3.137589043169972|
+-----+-----+
```

La distribucion temporal – los datos estan bien distribuidos con un enfoque claro de los primeros años del siglo

Variabilidad de conteo de los pasajeros – probablemente se debe de factores estacionales, de eventos o de cambio de tendencias de viaje

La concentracion de datos alrededor de 2010 sugiere que cualquier tendencia o patron puede ser mas representativo de ese periodo especifico



Analisis de correlacion para ver de que manera estan relacionadas las variables

	Activity Period	Passenger Count \
Activity Period	1.000000	0.060311
Passenger Count	0.060311	1.000000
Adjusted Passenger Count	0.059336	0.999941
year	0.999940	0.060069

	Adjusted Passenger Count	year
Activity Period	0.059336	0.999940
Passenger Count	0.999941	0.060069
Adjusted Passenger Count	1.000000	0.059096
year	0.059096	1.000000

Correlaciones casi perfectas entre Passenger Count y Adjusted Passenger Count. Significa que estas dos variables son casi intercambiables entre si. Cualquier ajuste no va a alterar significativamente la otra. Casi lo mismo para Activity Period y Year

Correlaciones bajas en el conteo de pasajeros. Activity Period y Passenger Count (0.060311), Year y Passenger Count (0.060069), Activity Period y Adjusted Passenger Count (0.059336) y Year y Adjusted Passenger Count (0.059096) indican que ni Activity Period ni year tienen una relacion lineal fuerte con el numero de pasajeros ya sea ajustado o no. Esto podría sugerir que el volumen de pasajeros no está directamente relacionado con el tiempo en una tendencia lineal simple, o que cualquier tendencia temporal en el conteo de pasajeros es superada por otros factores no capturados en estas variables.

Conclusiones:

Independencia de las variables de conteo de pasajeros – cualquier analisis realizado con la variables Passenger Count y Adjusted Passenger Count sera aplicable a la otra

Influencia Temporal Limitada: La baja correlación entre el Activity Period/year y los conteos de pasajeros sugiere que otros factores, posiblemente relacionados con eventos específicos, condiciones meteorológicas, cambios en las rutas de vuelo o políticas de precios,



pueden tener un impacto más significativo en el número de pasajeros que simplemente el paso del tiempo.

Potencial para Análisis Futuros: Dada la limitada correlación entre el tiempo y los conteos de pasajeros, podría ser útil explorar otros conjuntos de datos o variables que capturen eventos específicos, cambios en la industria de la aviación o patrones estacionales para comprender mejor las variaciones en el número de pasajeros.

Se aplica el algoritmo de regresion lineal por su simplicidad y eficiencia

```
RMSE: 561.1498617032486  
R2: 0.9999072855593121
```

RMSE indica el error promedio de las predicciones. La interpretación de este valor depende del rango y la escala de la variable "Passenger Count". Si el rango de Passenger Count es muy grande, entonces un RMSE de 561 podría considerarse bajo.

R2 tan cercano a 1 indica que el modelo puede explicar casi toda la variabilidad de la variable dependiente (Passenger Count) por sus variables independientes. Esto sugiere que el modelo es muy preciso en este conjunto de datos.