




Cloud y Big Data

Proyecto Final: GRUPO B

Enero 2017

Celia Gil Rodríguez
Óscar González Jiménez
Manuel Oreja Valverde
Alex Pascua Piña
Diego Pastor Calvo

Índice

- 
- 01. Descripción del proyecto
 - 02. Tecnologías para desarrollo
 - 03. Análisis de resultados
 - 04. Experiencias
 - 05. Futuras mejoras



01

Descripción del proyecto

Descripción del proyecto

La finalidad de este proyecto es ofrecer información necesaria para realizar una mejor planificación de sus vuelos a las aerolíneas o información al cliente para que decida qué vuelo le conviene mejor escoger.

Para ello, nuestra aplicación es capaz de analizar determinados parámetros de datos históricos, como puede cancelaciones o el retraso del vuelo y mostrar información procesada para realizar predicciones. De esta manera, la aerolínea puede anticiparse y replanificar horarios, o el cliente comprar un vuelo con una probabilidad baja de retraso de una cierta compañía.

El rango de los datos será de 3 años, manejando de esta manera información de aproximadamente 16.000.000 vuelos.

A vertical lighthouse stands against a dark, starry night sky. A bright beam of light emanates from the lantern room, extending towards the upper right corner of the frame. The lighthouse is white with a dark top section. The beam of light has a soft, ethereal glow.

02

Tecnologías para el desarrollo

Tecnologías

- Spark
- Python
- Capa SQL sobre Spark
- AWS (pruebas y desarrollo)
- Github: con el siguiente enlace para la visualización :
<https://aerodelays.github.io/>



03

Análisis de los resultados

Análisis de los resultados

Consultas en m4.large local (8 GB RAM):
45-50 segundos cada consulta

Consultas en m4.large cluster (8 GB RAM):
45-50 segundos cada consulta

Consultas en m4.xlarge local (16 GB RAM):
14-15 segundos cada consulta

Conclusión: Mejor escalar verticalmente. No está preparado para paralelizar las consultas.

04

Experiencias



Experiencias

Positivas:

- Hemos ganado experiencia con spark, además hemos aprendido parte de python.
- Gran experiencia en el trabajo en equipo y comunicación de resultados con vías telemáticas.
- Nos ha gustado mucho la experiencia de poder lanzar sentencias sql sobre spark.
- Poder comparar nuestros resultados con el de Google travels.

Negativas:

- El estudio de los fallos SQL nos ha llevado más tiempo que los de spark
- Nos ha fallado la organización, ya que las últimas semanas han sido mucho más agobiantes.


Experiencias

Ejemplo (Google Vs AeroDelays):

Vuelo de ida

X

jue., 1 feb.

 **07:30 – 15:26**
Seattle (SEA) - Washington D. C. (IAD)
United 419 · Boeing 737
Espacio normal para las piernas (76 cm)

4 h 56 min
Wi-Fi
Conexión en asiento
Transmisión de vídeo a dispositivos personales

```
.....+  
avg(ARR_DELAY_NEW) |  
.....+  
21.179640718562876 |  
.....+
```

A close-up photograph of a person's hands interacting with a tablet device. The image is heavily stylized with a green translucent overlay that covers most of the frame. The hands are visible, with fingers touching the screen. In the background, a blurred computer monitor is visible. The overall aesthetic is modern and technological.

05

Futuras mejoras

Futuras mejoras

- Parsear los resultados obtenidos.
- Ampliación del rango de datos, cogiendo los datos desde 1987.
- Hacer más consultas, más complejas para la obtención de resultados más precisos.
- Unir a interface para no mostrar la ejecución de spark.
- Utilizar patrones estadísticos robustos para realizar mejores predicciones.
- Paralelizar las consultas para clusters.

Futuras mejoras

Ejemplos:

Archivo Editar Ver Buscar Terminal Ayuda

Selecciona una opción

opción 0: Media de retraso
opción 1: Probabilidad de retraso
opción 2: Probabilidad de cancelacion
opción 3: Probabilidad de cancelacion por causa
opción 4: Media de retraso en salida
opción 5: Probabilidad de retraso en salida
opción 6: Media de retraso en salida (solo vuelos retrasados)
opción 7: Media de retraso en llegada
opción 8: Probabilidad de retraso en llegada
opción 9: Media de retraso en llegada (solo vuelos retrasados)
opción 10: Media de retraso por aerolinea
opción 11: Media de retraso por aerolinea (solo vuelos retrasados)
opción 12: Probabilidad de retraso por aerolinea
opción 13: Probabilidad de cancelacion por aerolinea

3
Escribe el origen:
31057
Escribe el destino:
31136

```
+-----+
| CANCELLATION_CODE | (CAST(count(1) AS DOUBLE) / CAST(scalarsubquery() AS DOUBLE)) |
+-----+
| B | 0.8851674641148325 |
| C | 0.028708133971291867 |
| A | 0.0861244019138756 |
+-----+
```