CASO PRÁCTICO 1

Se han recolectado los datos de una compañía aérea de los vuelos realizados en 25 países durante el mes de febrero. Se dispone de 4 archivos:

1. **países**: relaciona código de país con el País correspondiente
2. **vuelos**: recopila información sobre el número de vuelo, origen y destino
3. **retrasos**: información sobre el retraso que ha tenido el vuelo (valores entre 10 y 99 minutos)
4. **fecha**: día de febrero que tuvo lugar el vuelo

**** El archivo airTribu.xlsx, es un Excel con el resumen de todos los archivos.***

Primeros pasos:

- Subir los archivos a HDFS.
- Crear tablas en HIVE relacionando los archivos.

Preguntas:

1. ¿De qué país salieron más aviones?
2. ¿A qué país llegaron más aviones?
3. ¿Qué día hubo más y menos vuelos?
4. ¿Qué día hubo más y menos retrasos?
5. Crear un tabla resultado que tenga la información del origen de los vuelos y su retraso acumulado por día (sin importar el destino). Ejemplo, si partimos de los siguientes datos:

origen	**destino**	**día**	**retraso**
AAA	XXX	1	5
BBB	YYY	3	10
AAA	BBB	2	20
BBB	QQQ	20	15

el resultado sería:

origen	**día**	**retraso**	**retraso_acumulado**
AAA	1	5	5
AAA	2	20	25
BBB	3	10	10
BBB	20	20	30

Pista: Window function

6. Sobre el resultado del ejercicio 4, añade otra columna que sea **"Pais_VIP"** donde se identifique si un país es VIP (los países VIP son España, Perú y México tanto en origen como destino), para ello haz uso de UDF
7. Si se desea almacenar la información del resultado del ejercicio 4 en solo 1 archivo, ¿cómo lo harías? ¿y si lo quisiera en 10?
8. Sobre el resultado del ejercicio 6, salva en una tabla **Hive** solo cuando el origen sea Perú. Después de eso vuelve a salvar sobre la misma tabla el resultado de filtrar cuando el origen es México. Por lo tanto, en la tabla **Hive** resultado de este ejercicio deben aparecer los registros de países Perú y México.
9. ¿Cómo escribir un **dataframe** con particiones de fecha?, dar ejemplo de uso
10. ¿Qué es el plan lógico de **spark** y como optimizar querys?
11. ¿Cómo usar el **Spark UI?**, dar ejemplo de uso
12. ¿Qué es skew data, y como superar este problema en Spark?
13. ¿Cuál es la diferencia entre **cache** y **persist?**, dar ejemplo de uso
14. Explica que es el efecto **shuffle** y cómo afecta al procesamiento de grandes volúmenes de datos.
15. ¿Cómo identificar el **shuffle** en un **dataframe** y cómo corregirlo?
16. ¿Qué es **bucketing**?, dar ejemplo de uso

PRESENTACIÓN:

- **Resumen ejecutivo (PPT) – Templates**
https://www.indrabrandcenter.com/document/65#/plantillas-office/power-point

- Código fuente del desarrollo:

*** Desarrollo con Clean Code**

*** Subir a un repositorio GIT**