**Pregunta 1.** Tras analizar el dataset de datos aeroportuarios. ¿Qué es lo que más te ha llamado la atención? ¿Qué datos te hubiese gustado que incluyese?

**Respuesta 1.** Lo que mas me ha llamado la atención es que la mayoría de datos son categóricos. Cuando se habla de BigData yo siempre he tenido en mente muchos números, pero no es el caso de este DataSet.

En general el DataSet me parece bastante completo y permite hacer un buen análisis del tráfico aéreo de nuestro aeropuerto.

Se podrían añadir datos más concretos sobre nuestros pasajeros como el sexo o la edad. Estudiar estos datos podría servir para adecuar los servicios que ofrecemos en nuestro aeropuerto.

Se podrían añadir también datos sobre las horas de salida y llegada de los vuelos para estudiar los momentos en los que más pasajeros eligen volar (madrugada, mañana, tarde, noche…).

**Pregunta 2.** En el apartado de Spark has utilizado comandos de sql para la limpieza de los datos. ¿De qué otra forma podrías haber utilizado que no fuesen sentencias de sql?

**Respuesta 2.** Podríamos haber utilizado la API propia de PySpark (código en google collab, parte 2, apartado2)

**Pregunta 3.** Explica cómo funciona una Primary Key en Cassandra.

¿Podrías explicar por qué eliges los campos "activity\_period" y "operating\_airline" para generar la primary key?

¿Por qué usas operating\_airline como clustering key?

¿Cambiarías algo del trabajo presentado?

**Respuesta 3. La row key o clave primaria** consta de tres elementos:

Partition key: Se usa para identificar el nodo o servidor donde se va a almacenar el dato. Todos los datos que posean la misma partition key se ubicarán en la misma partición.

Static columns: Son columnas estáticas que no modifican su valor dentro del mismo partition key. De esta forma, si dos filas poseen el mismo partition key, han de presentar el mismo valor que las columnas estáticas.

Clustering key: Sirven para indicar el orden en el que se almacenan los datos.

Establecemos activity\_period and operating\_airline como partition key para distribuir uniformemente los datos en las particiones. Es cierto que los datos de ambos campos provocan muchas particiones relativamente pequeñas, pero dado que el número de particiones que puede soportar cassandra es muy elevado (2^63) es preferible tener un gran numero de particiones pequeñas para evitar que estás, en un futuro, si se añaden datos, crezcan demasiado. Mejor mantener las particiones pqueñas

Dado que cada partición va a contener los datos, únicamente, de un periodo y aerolínea concretos, la elección de operatin\_airline como clustering key es poco acertada ya que la función de la cluster key es el optimizar la recuperación de datos en un orden concreto. Explicado esto, una elección mas adecuada habría sido utilizar otro campo que no se encontrara en la partition\_key, como passenger\_count.

Esto sería definitivamente, el cambio que realizaría en el trabajo.