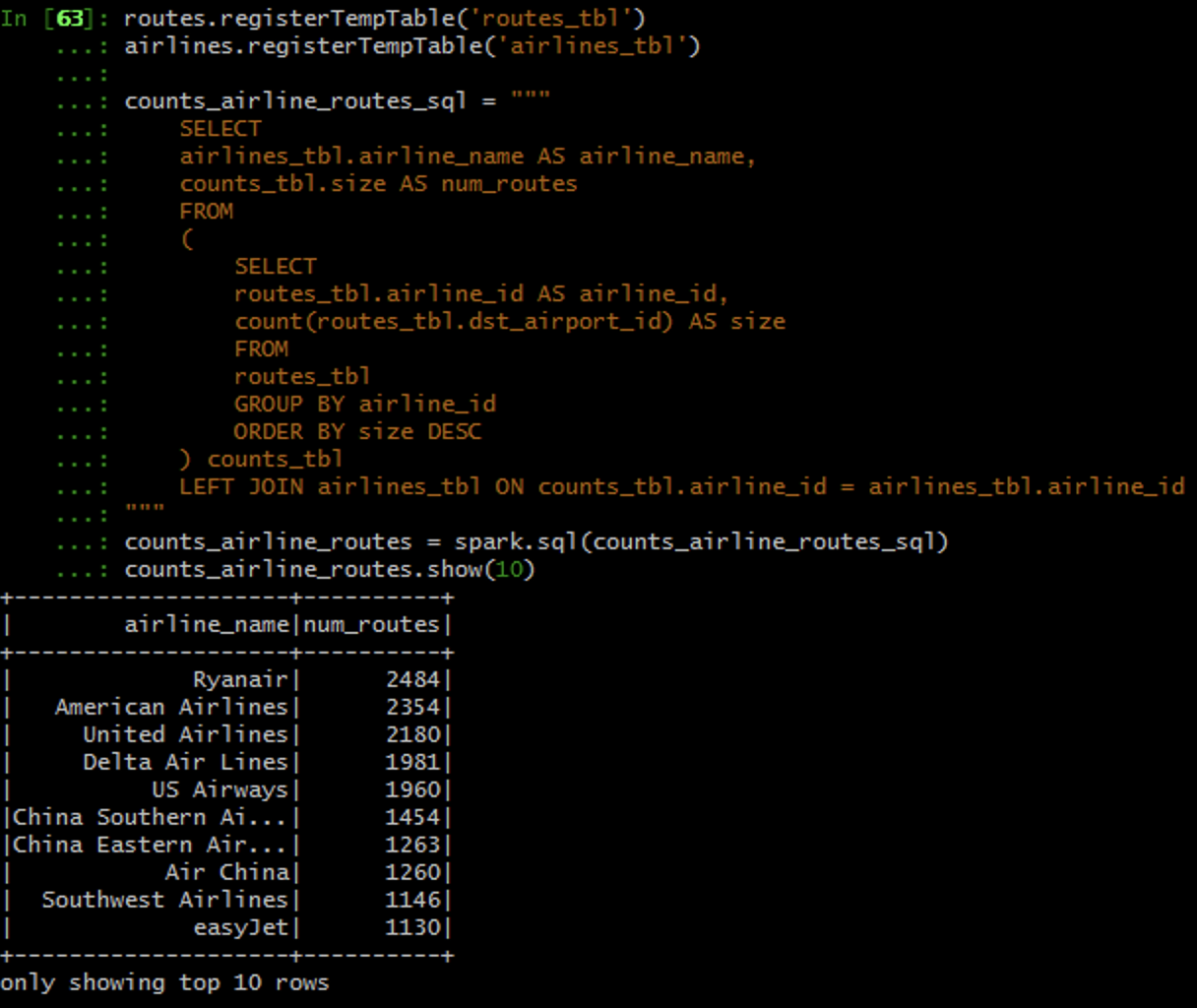
***Q1 (a)***

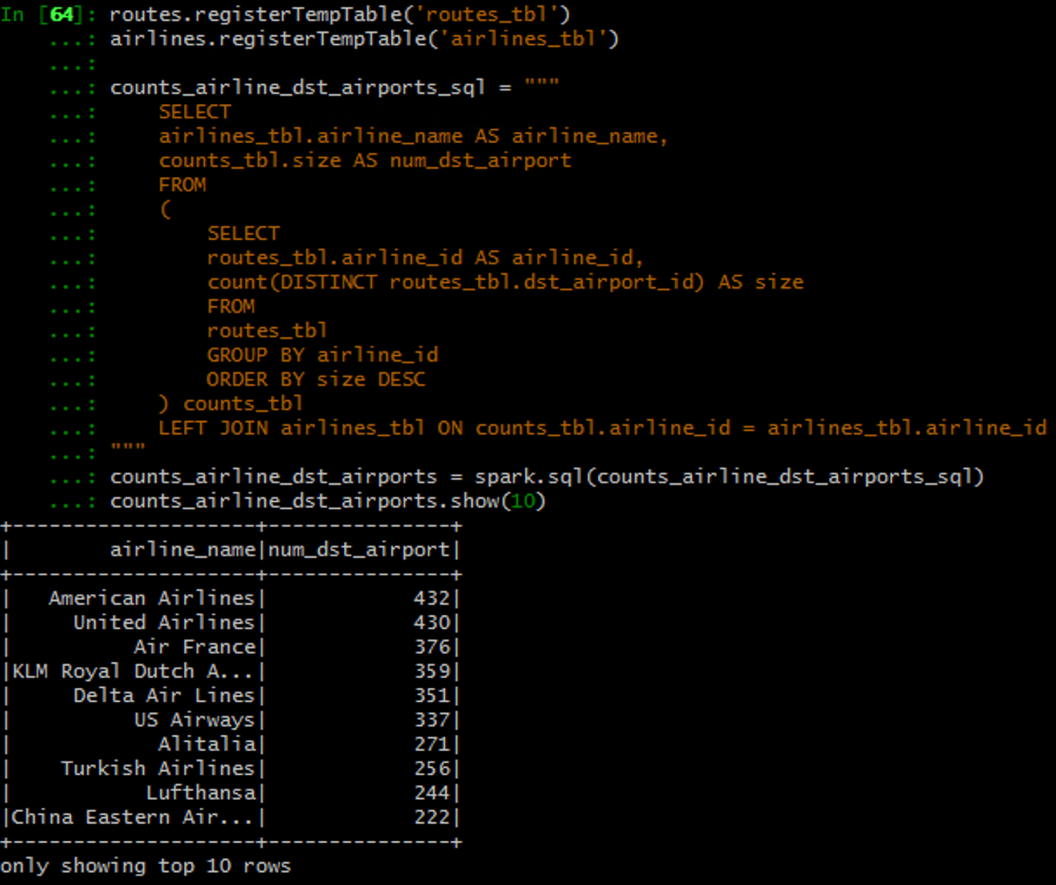
*Describe in your own words the data transformation that each step is doing. Is there any difference between the data transformation done by Spark SQL and the Python Spark API?*

*Line 18 - 64 in the code above defines the schema and datatype for the RDD of routes and airports. The line 68-69 tranform the RDD into a table that can be manipulated using Spark SQL . Line 71-88 using Spark SQL create a dataframe showing each airport\_name and the number of airports connected to it. Line 92 - 122 using Pyspark SQL API to do the same task above. Spark SQL need to transform the RDD into table to process, then do the task by joining tables, selecting columns, counting, grouping, ording, sometimes under certain conditions. Pyspark SQL API can deal with RDD directly without the transformation above, besides, the whole workflow can be devided into branches, ie., by defining and using components to make it easier to understand and develop.*

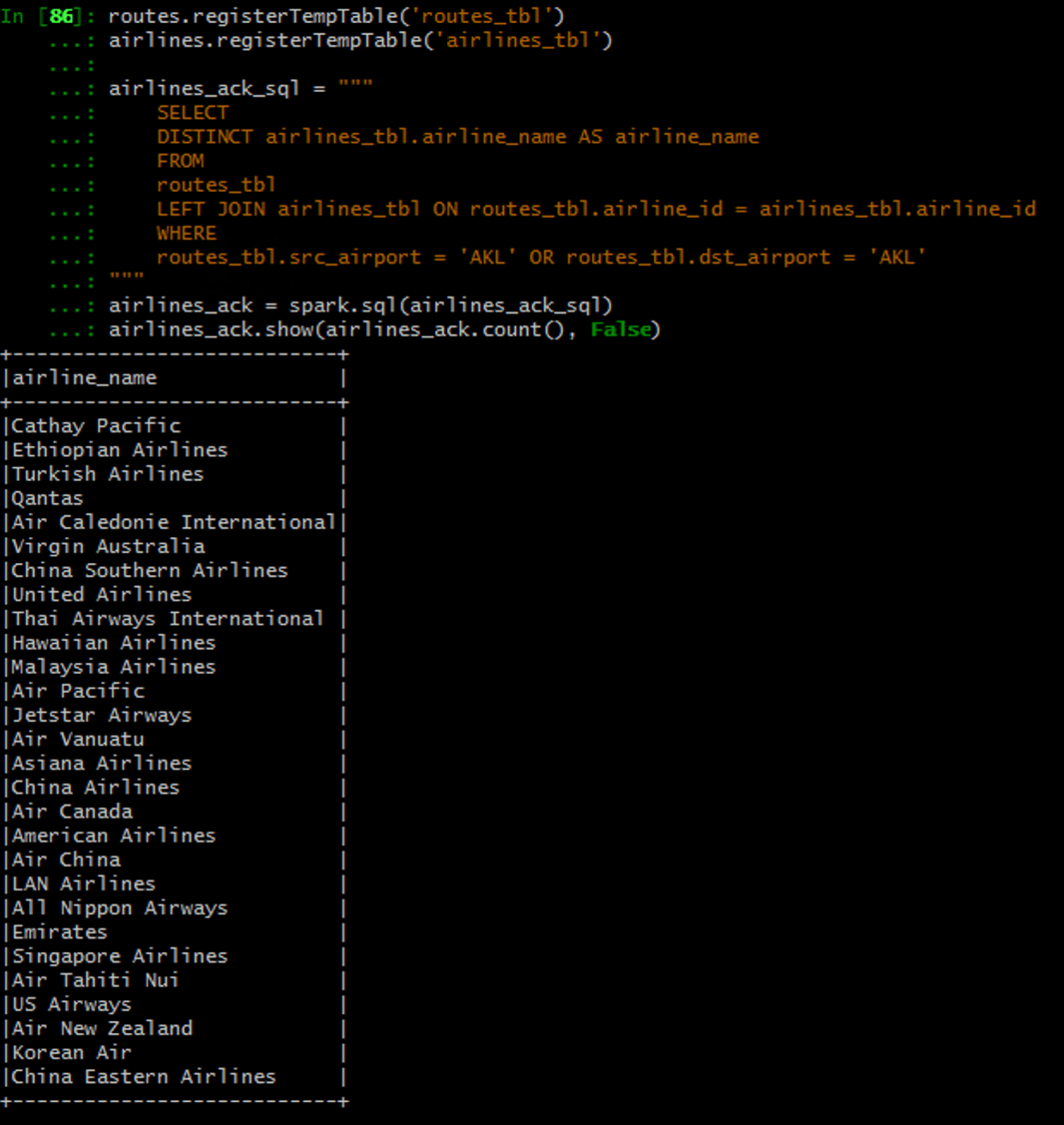
*Q1 (b)*

**

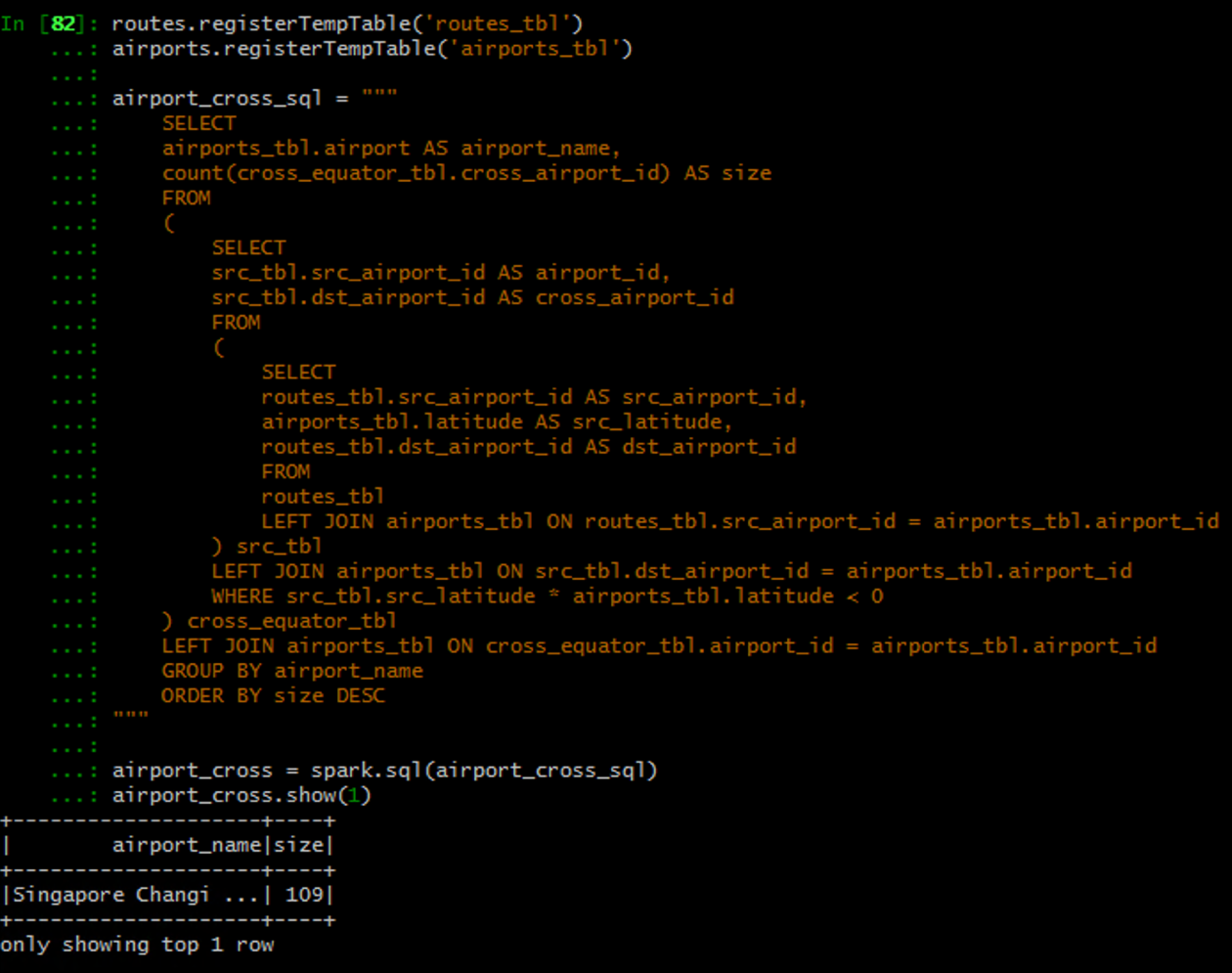
*Q1 (c)*

**

*Q1 (d)*

**

*Q1 (e)*

**

