Challenges faced in Spark 301 – Air Routes

1. Directly we were not able to load the data to RC format in Hive hence we created the temporary tables.

2. Air Routes datasets had direct datasets and tried to load them, then realized the data was not loaded properly. There was a data format issue .Also found there are double quotes inside double quote creating this problem. Thus we wanted to cleanse data and so we came up with below solutions

* Clean data using shell script
* Clean data using Pig script
* Clean data using CSV SERDE file in Hive itself

Thus we took the effective ones i.e., clean data using CSV SERDE file while creating a temp table in Hive itself and we did not have to use any other external tools for this

3.In Intellij community edition, we were getting an error (Scala class path not found) while creating a jar file. Then we found we have to select the lower Scala version than the SBT version

4.In Intellij, though we could create a jar file successfully, we couldn’t execute it as there was an exception (Java class not found exception), thus we selected “Include in Project build” while building the artifacts

5Cassandra tables won’t be mapped if the selected hive table columns aren’t given in lowercase (ex: airline\_id), thus we changed the column names to lowercase in SparkSQL queries