Problem 2

Writing custom Input Format to read multi-line Json file

The solution contain 3 java scripts

Query2.java driver

JsonRecordReader.java provide method to read multiline Json file

JsonInputFormat.java will consume JsonRecordReader and be used in the driver

Explanation:

We created a new input format called JsonInputFormat, which will use newly created class JsonRecordReader. In the next method of JsonRecordReader, we define when the reader come across “}” or “},”, the reader will stop for next record, otherwise, it will include all the content into one record.

The JsonInputFormat.java is later used by Query2 in job configuration job.setInputFormatClass

JsonRecordReader will emit key-value pair in the format of LongWritable and Text.

In the Query2.java, we have MRMapper class, where flag number is extracted by substring function as a key and counter is included for each record as a value. In the MRReducer class, counter is aggregated based on flag number.

For dividing the file into 5 split, we need to tweak the size of maxium split size. To do that, in JsonInputFormat, we override the computeSplitSize method, I was able to set it use 3 mapper , it works fine. When we set the size use 5 mapper, it has an “java.lang.ArrayIndexOutOfBoundsException: 5” error. Likely, it is due to visual machine setting.

How to run

To run this in Hadoop, you need to export the project into an executable jar.

Save the jar file to the Hadoop local folder

At linux terminal, change prompt directory to folder where your jar is saved

Use the following command

Hadoop jar jarname airfield.txt hdfs path output hdfs path

For example

Hadoop jar Q2.jar /tmp/airfield.txt /tmp/output40