Changes between old A4 and current A4

Bishwajeet Dey

November 11, 2017

Architecture and Design

- The old A4 had 2 map reduce jobs. The first job did the cleaning and the second job calculated the top 5 airports and airlines. The second job was not required since it processes only 11 MB of text. The new design only has one job which does the cleaning and produces the aggregated output.
- The old cleanup job did not correctly compare 24 hour time for departure and arrival time. This led to incorrect number of active airports and airlines. This was also fixed in the current version.
- There were subtle bugs in the data validation where we had used a regex for data validation. It was replaced with simpler string validation checks. Overall, large parts of the code were removed or re-written.
- The old project had 22 java files. Most of the files were unused. The current project has 10 java files and 1 scala file for spark.
- The old design was not explained properly. This led to a lot of confusion from the reviewers about the flow of code. The current report explains it in the beginning.
- Finally, the current project also features spark code apart from the Hadoop framework code.

Analysis

- The old project reported incorrect active airports and airlines due to a bug in cleanup. This was fixed in the current version.
- In the previous report, we concentrated on seasonal analysis only for the years 2004-2009. This led to missing the trend of delays over the past 30 years. The current report focusses on the trends and analysis for the whole time period. It also does seasonal analysis which has more insights than the previous report.
- The previous report lacked analysis of the graphs. We could not see the trend in delays for airports and airlines.
- The graphs and analysis were not formatted correctly. This led to graphs being split across pages and missing legends. The graphs in the current report fit within a page.