Report Project 2: BDAP [B-KUL-H00Y4A]

Andreas Hinderyckx

April 2022

Part 1: Trip Length Distribution

Part 2: Computing Airport Revenue

Identifying erroneous GPS points The first technique used to identify erroneous GPS points is the one suggested in the assignment: the calculated speed of a segment must be smaller than 200 km/h. A lower speed limit was not chosen, as GPS results may be imprecise at times, which could cause valid but slightly inaccurate trips to be skipped as well. Secondly, it was apparent that quite some GPS points were located in the sea, to the West of San Francisco. On Google Maps, two coordinates were handpicked to create a rough approximation of the coastline. Points located to the west of this approximated coastline were rejected. This check rejected about 1500 points, but did hurt performance however (see below). Finally, between every two segments of a trip, it is checked whether the elapsed time between the end time of the first segment and the start time of the second segment doesn't exceed three seconds.

Reconstructing the trips The trips are reconstructed in the first of the two map-reduce-processes. Firstly, each segment is given a key: `TaxiID,StartDate`: this makes sure that the segments are primarily sorted on `TaxiID` and secondary sorted on `StartDate`. The `mapper` also rejects malformed records (see below) and doesn't map `E-E`-records as these don't contribute to a trip. Additionally, a `Partitioner` is added to ensure that segments with the same `TaxiID` end up at the same reduce-task.

Next, given the sorted order of the records, the trips can be constructed during the reduce-phase. The reducer scans for a records that transitions from the `E` state to the `M` state to initiate a trip. subsequent records are added to the trip, until a records is encountered that transitions from the `M` state to the `E` state. During this process, the reducer rejects trips that contain a records with excessive speed, or that contain records with a time gap in-between them that is larger than 3 seconds.

Efficiency of the solution TODO Due to the check that verifies whether a GPS point is not located at sea, the execution time of the program rose with 30% (from $\sim$ 9 to $\sim$ 12 minutes). As these mislocated records only constitute $5 \cdot 10^{-4}\%$ of the total amount of records, this is a significant performance penalty to pay for a small gain in accuracy.

Total revenue & revenue over time