CPSC 5330 01 - Big Data Analytics
Lab 02
Airport Delay Analysis using Hadoop Streaming

Fariha Shah

April 20, 2025

1 How much time you spent on the assignment

I spent approximately 12–14 hours on this assignment. Time was distributed across understanding the data structure, writing the mapper and reducer scripts, setting up the Hadoop Streaming job, and debugging environment-related issues.

2 If parts of the assignments are not fully working, which parts and what the problem(s) are

- Initially, my script failed due to incorrect path references to the Hadoop streaming JAR file. This was resolved by locating the correct path manually using the find command.
- Output wasn't being generated at first (/airport-delay/output/part-* not found), which was also due to the streaming job not running successfully — again linked to the JAR path.
- While the final job produced results, additional validation and formatting may still be needed for very large datasets or inconsistent JSON structures.

3 Were there aspects of the assignment that were particularly challenging? Particularly confusing?

- One major challenge was identifying the correct Hadoop streaming JAR location. This caused multiple script runs to fail and delayed progress.
- Another tricky part was handling malformed or incomplete JSON records. These needed to be safely ignored in the mapper.py using try-except blocks.
- Understanding that each JSON file represented one month at a single airport was initially unclear, which could affect aggregation if not handled properly.

4 What were the main learning take-aways from this lab – that is, did it introduce particular concepts or techniques that might help you as an analyst or engineer in the future?

- I gained hands-on experience with Hadoop Streaming, which reinforced how to use custom Python mappers and reducers.
- Learned how to safely parse JSON and handle unexpected edge cases in a distributed environment.
- Understood the process of using HDFS for data upload and output, and how to connect Sqoop, Hadoop, and shell scripting into a cohesive data pipeline.
- Gained experience writing end-to-end scripts (report-airport-delays) that automate a complete job cycle from data ingestion to output generation.
- Realized the importance of careful environment setup, especially when working with distributed systems like Hadoop.