Log Extraction Project

Project Context

Live services generate logs at a very high rate; e.g., our service creates over 100,000 log lines a second. Usually, these logs are loaded inside databases to enable fast querying, but the cost of keeping all the logs becomes too high. For this reason, only the recent logs are kept in databases, and logs for longer periods are kept in file archives.

For this problem, we should assume we store our data in multiple files. We close a file and start a new file when the file size reaches 16GB. Our file names are of the format LogFile-######.log (e.g., LogFile-000008.log, or LogFile-000139.log). Currently, we have over 10,000 log files with the last log file name LogFile-0018203.log and a total data size of 285TB.

Problem Statement

We usually use our log database to query our logs. But now and then, we may have to query older logs for customer support or debugging. In most of these cases, we know the time range for which we need to analyze the logs.

We need a tool that could extract the log lines from a given time range and print it to console in time effective manner.

The command line (CLI) for the desired program is as below

```
LogExtractor.exe -f "From Time" -t "To Time" -i "Log file directory location"
```

All the time formats will be "ISO 8601" format.

The extraction process should complete in a few seconds, minimizing the engineer's wait time.

Log file format

1. The log file has one log per line.
2. Every log line will start with TimeStamp in "ISO 8601" format followed by a comma (',').
3. All the log lines will be separated by a single newline character '\n'.
4. Example logline:
   ```
   2020-01-31T20:12:38.1234Z, Some Field, Other Field, And so on, Till new line,...\n
   ```