Report for Data Deduplication Thesis

In this report i have did performance analysis and scalability and robustness. My datasets were made out of 1MB(for scalability and robustness), 50MB, 100MB, 500MB and 1000MB(for performance).

The Test Environment:

* Samsung SSD 980 PRO 2TB (M2.SSD)
* Intel Core I7-13700K 13th generation CPU
* 64GB 3200MHz Ram
* NVIDIA Geforce RTX 4070 Ti GPU
* Windows 11 Pro 64-bit (10.0, distribution 22631)

Performance Tests:

In this test I had 4 separate data sets which had 10, 50, 100, 250 and 500 files which had 5 sub-datasets.

50MB files:

|  |  |  |
| --- | --- | --- |
| Number of Files | Scan Time(s) | Compare time (s) |
| 10 | 2.00 | 0.02 |
| 50 | 10.77 | 0.33 |
| 100 | 29.53 | 1.31 |
| 250 | 54.21 | 8.77 |
| 500 | 109.49 | 38.24 |

100MB Files:

|  |  |  |
| --- | --- | --- |
| Number of Files | Scan Time(s) | Compare time (s) |
| 10 | 3.75 | 0.02 |
| 50 | 18.76 | 0.33 |
| 100 | 37.93 | 1.37 |
| 250 | 106.85 | 8.60 |
| 500 | 218.61 | 40.30 |

200MB:

|  |  |  |
| --- | --- | --- |
| Number of Files | Scan Time(s) | Compare time (s) |
| 10 | 20.36 | 0.02 |
| 50 | 101.85 | 0.31 |
| 100 | 208.43 | 1.29 |
| 250 | 530.35 | 8.66 |
| 500 | 1084.11 | 39.24 |

1000MB:

|  |  |
| --- | --- |
| Number of Files | Scan Time(s) |
| 10 | 41.71 |
| 50 | 209.07 |
| 100 | 424.50 |
| 250 | 1049.63 |

Comparing the Results:

In this graph we observe that as the single size of the data grows the time it takes to scan it grows too. With this data we can observe the time complexity using polynomial regression.

50MB Data:

100MB Data:

500MB Data:

1000MB Data:

Interpretation:

With these models we observed that The linear term(n) is significant in all cases, indicating a linear relationship with the number of folders. And The quadratic term(n^2) is relatively small but positive in the 100MB and 500MB data, suggesting a minor quadratic influence. For the 50MB and 1000MB datasets, the quadratic term is negative but very small, indicating the dominant effect is linear.

Time Complexity:

From the coefficients, the linear term is dominant across all datasets, indicating that the time complexity is approximately O(n). This indicates that the system performs linearly with respect to the number of folders, even as the dataset size increases. The minor quadratic terms suggest slight deviations, but they are not significant enough to alter the overall linear trend.

Scalability and Database Performance Tests:

In this dataset I had 1 Dataset of 1MB files with 7 sub-datasets which were: 100, 1000, 2000, 5000, 10000, 20000 and 50000 files. And with this test I observed the data scan times and database select operation times.

|  |  |  |
| --- | --- | --- |
| Number of Files | Scan Time(s) | SELECT Time(s) |
| 100 | 1.41 | 0.0060 |
| 1000 | 17.96 | 0.0099 |
| 2000 | 35.80 | 0.0149 |
| 5000 | 91.81 | 0.0213 |
| 10000 | 178.97 | 0.0364 |
| 20000 | 361.88 | 0.0799 |
| 50000 | 1039.89 | 0.1678 |

After applying polynomial regression models to these data sets, we observe the following models:

Scanning:

We have observed that the time complexity is O(n).

Database SELECT time:

We have observed that the time complexity is O(n). But the time it takes to select the data from the database is so small overall, it is negligible.

The scalability analysis demonstrates that both the scanning and database select operations have linear time complexity, making the system scalable with respect to the number of files. The minor quadratic terms are negligible, affirming that the dominant time complexity is linear for both operations.

This indicates that the system should handle increasing numbers of files efficiently, maintaining performance within an acceptable range as the dataset grows. ​