Benchmark Compression with Near-Sortedness

By -
Shivangi
Vani Singhal

Background

Why need data compression?

Large amount of data generated

Run out of resources soon

Motivation

Degree of Sortedness

Completely Sorted or Unsorted Data

Problem Statement

Explore the Performance

Workload

Different size of Workload

Different values of K-L
(Varying Sortedness)

5 Compression Algorithms

Run Length Encoding

Delta Encoding

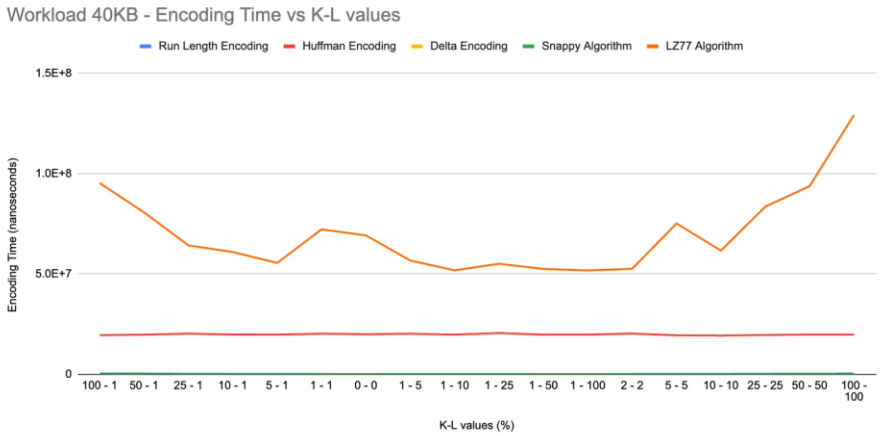
Snappy

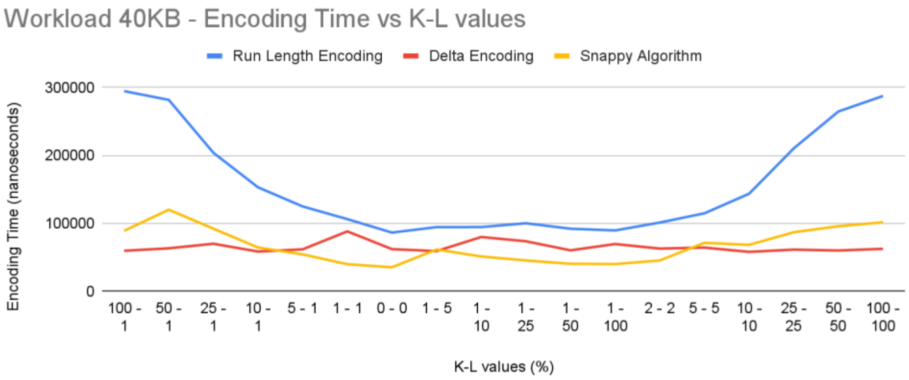
LZ77

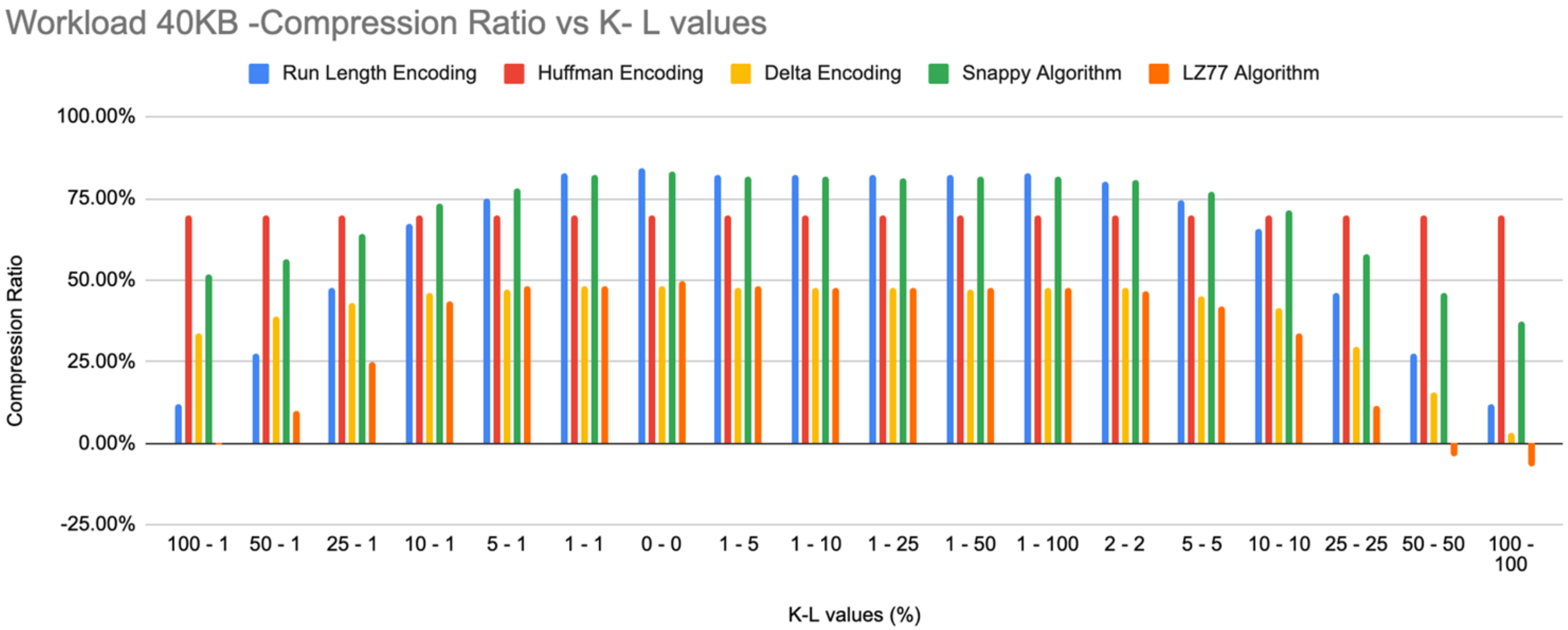
Huffman

Results

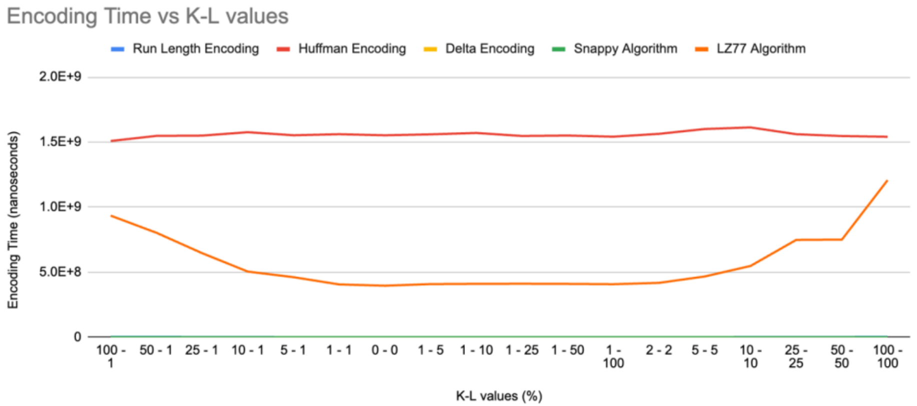
Workload 40KB

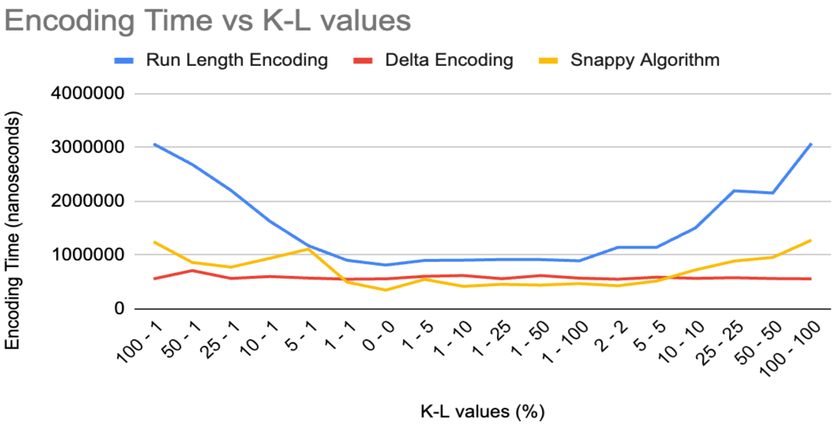


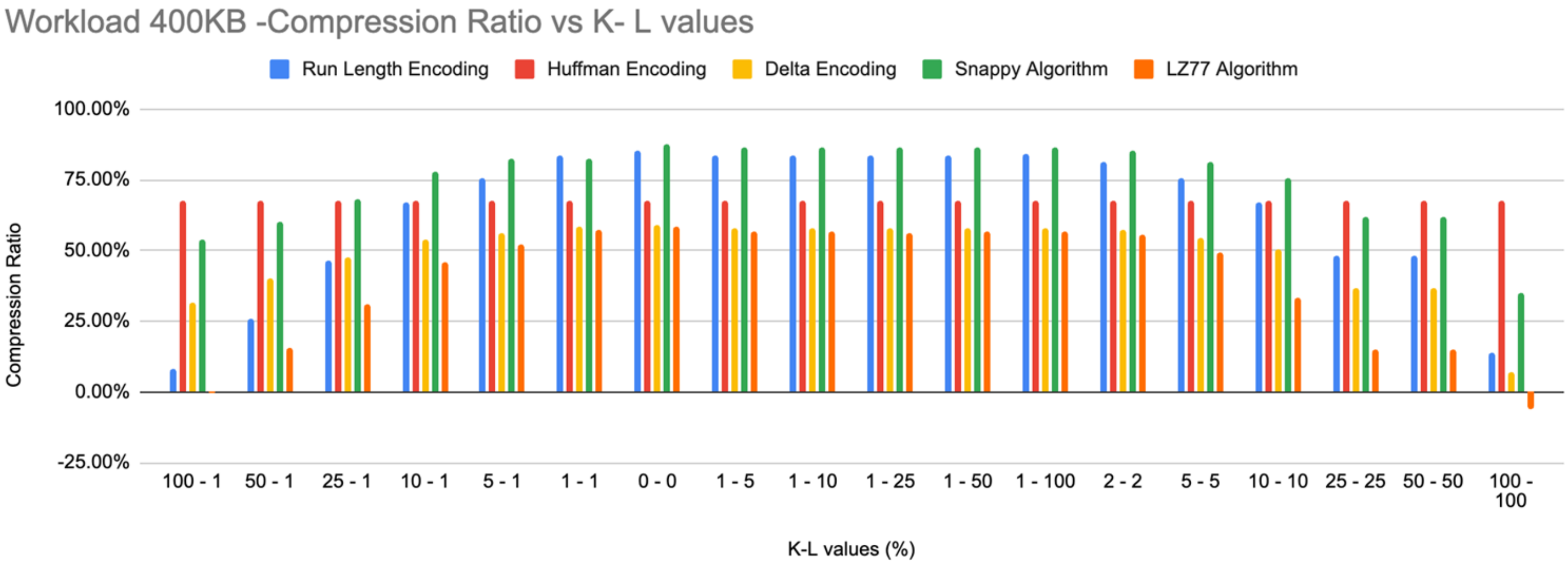



Workload 40KB -Compression Ratio vs K- L values
Run Length Encoding
Huffman Encoding
Delta Encoding
Snappy Algorithm
LZ77 Algorithm
100.00%
75.00%
50.00%
25.00%
0.00%
-25.00%
Compression Ratio
100 - 1
50 - 1
25 - 1
10 - 1
5 - 1
1 - 1
0 - 0
1 - 5
1 - 10
1 - 25
1 - 50
1 - 100
2 - 2
5 - 5
10 - 10
25 - 25
50 - 50
100 - 100
K-L values (%)

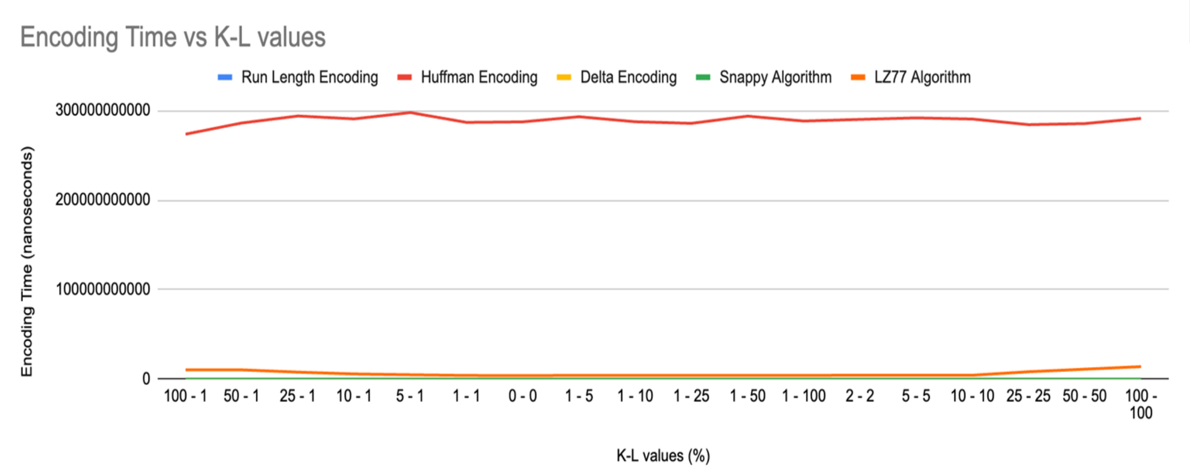
Workload 400KB

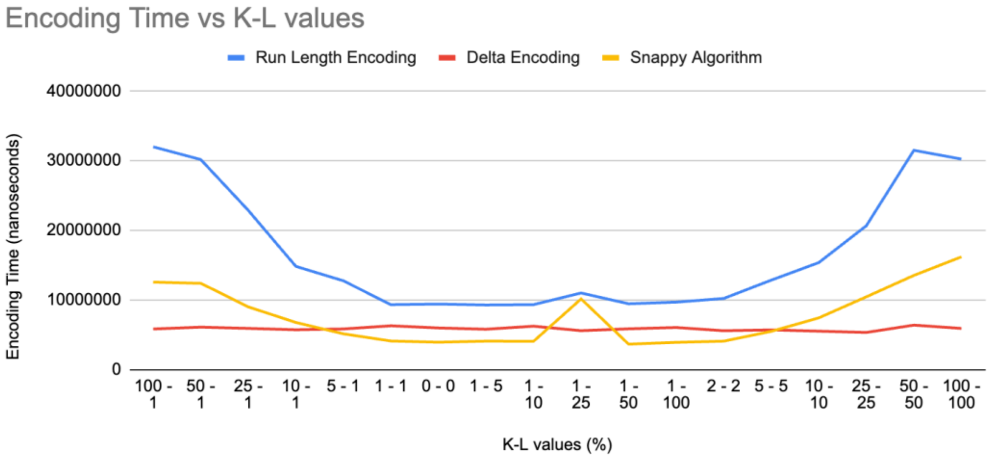


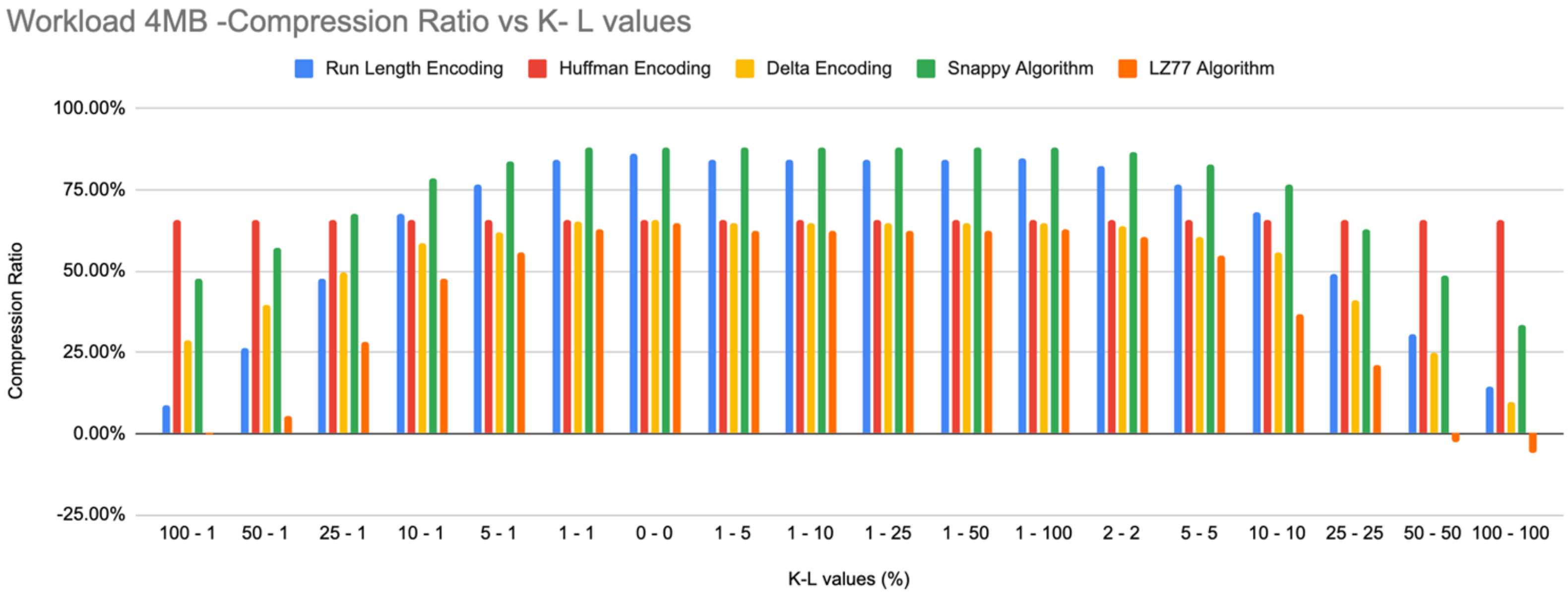



Workload 400KB -Compression Ratio vs K- L values
🟦 Run Length Encoding
🟥 Huffman Encoding
🟨 Delta Encoding
🟩 Snappy Algorithm
🟧 LZ77 Algorithm
Compression Ratio
100.00%
75.00%
50.00%
25.00%
0.00%
-25.00%
100 - 1
50 - 1
25 - 1
10 - 1
5 - 1
1 - 1
0 - 0
1 - 5
1 - 10
1 - 25
1 - 50
1 - 100
2 - 2
5 - 5
10 - 10
25 - 25
50 - 50
100 - 100
K-L values (%)

Workload 4MB

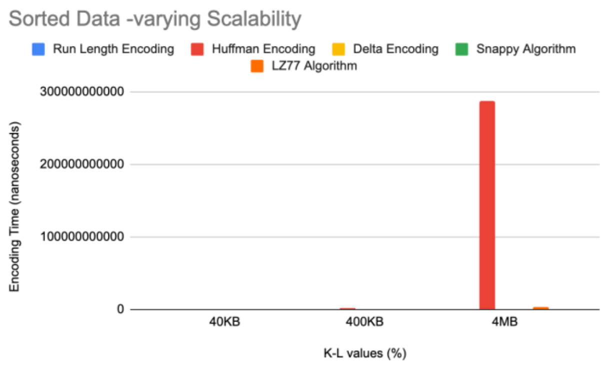


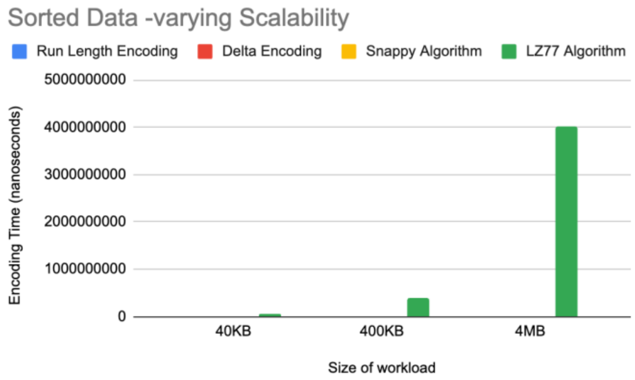


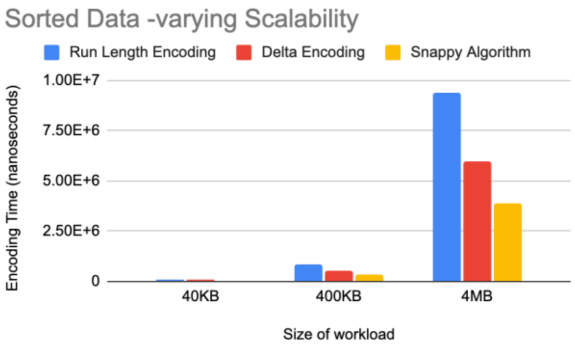


Workload 4MB -Compression Ratio vs K- L values

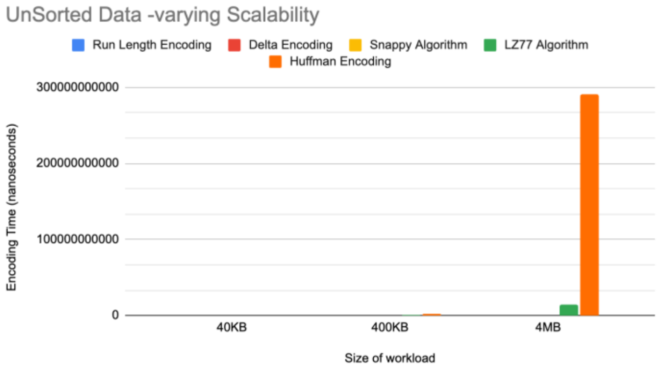
Scalability for Sorted Data

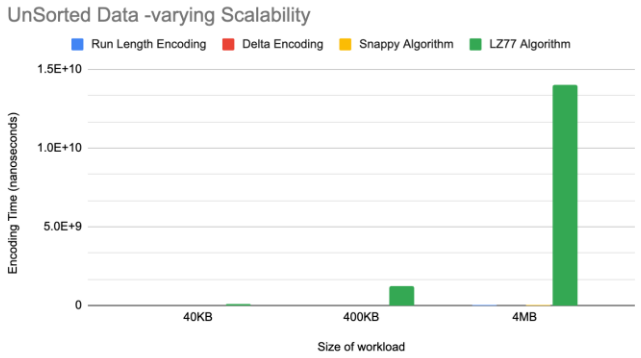


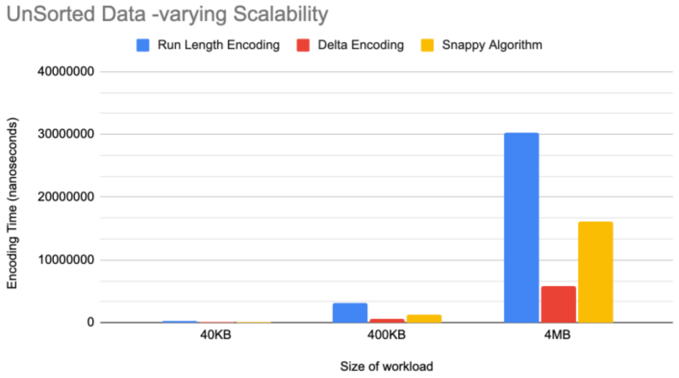




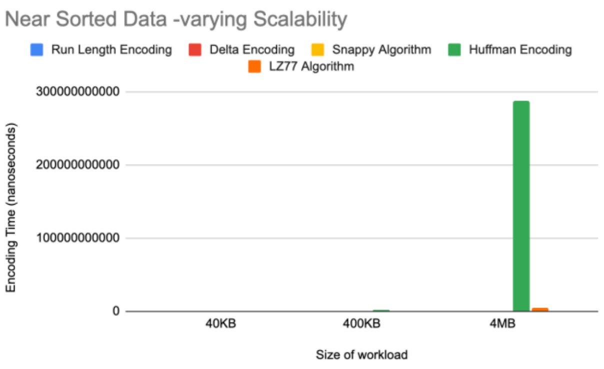
Scalability for Unsorted Data

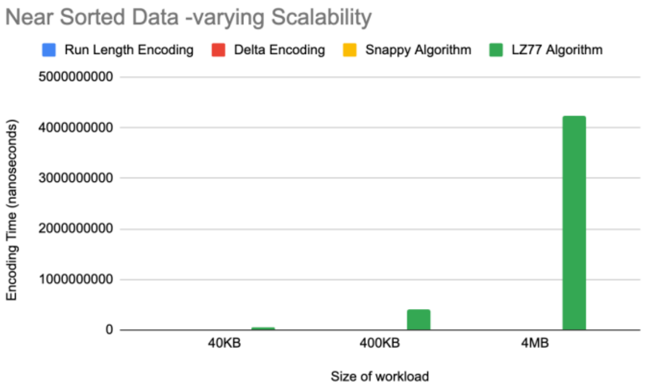


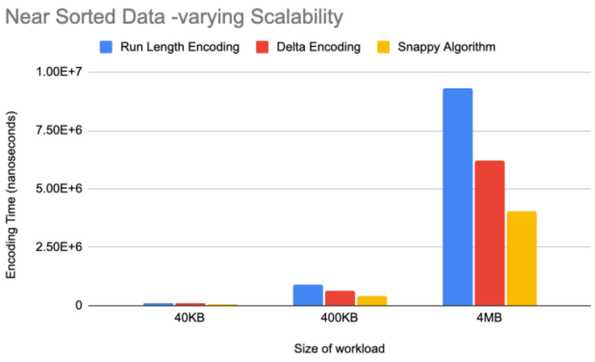




Scalability for Near Sorted Data







Next Steps

Unique Data

Larger Size of Workload - 40MB and 400MB

Questions ?