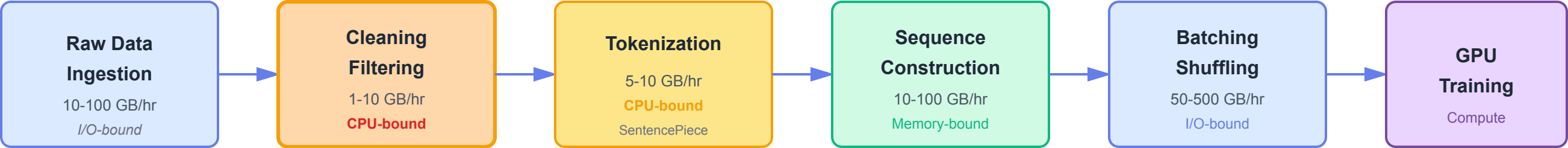


Data Preprocessing Pipeline Architecture



Pipeline Bottleneck Analysis

Disk I/O Bottleneck

SSD: 500-3,000 MB/s
HDD: 100-200 MB/s
Limits large datasets

Solutions:

- Parallel loading
- In-memory caching

CPU Processing Bottleneck

Single-thread: 1-10 GB/hr
Tokenization, filtering
GPU sits idle

Solutions:

- Multiprocessing (4-8 workers)
- Vectorized operations

Memory Bandwidth

PCIe 3.0: ~12 GB/s (CPU-GPU)
PCIe 4.0: ~24 GB/s (CPU-GPU)
Limits: Small batches, frequent transfers

Solutions:

- Larger batches (amortize overhead)
- Prefetching (overlap transfer)