In-Situ Processing of Genomic Sequence Alignment Data

Florin Rusu
Yu Cheng
Suzanne Sindi

UCMERCED

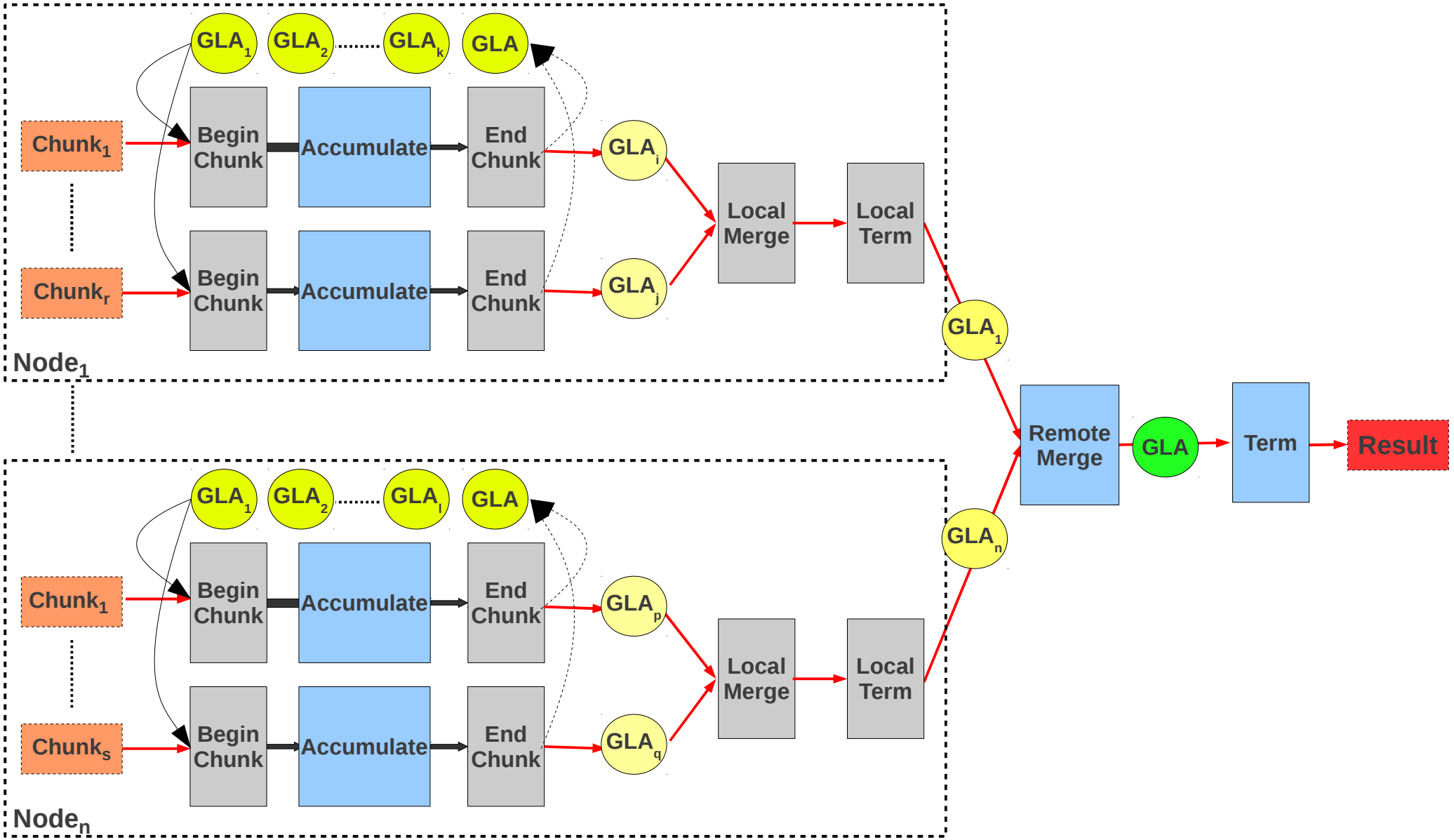
Sequence Alignment Data

- Output of sequencing machines is SAM/BAM files
  - SAM is tab-delimited text
  - BAM is compressed binary (compression ratio is 3-4)
  - Massive size (hundreds of GB)
  - Hundreds of millions of alignments (reads)
  - **Read** (QNAME:*string*, FLAG:*int*, RNAME:*string*, POS:*int*, MAPQ:*int*, CIGAR:*string*, RNEXT:*string*, PNEXT:*int*, TLEN:*int*, SEQ:*string*, QUAL:*string*)
- Tools: SAM/BAMtools, Picard, IGV, …
  - File manipulation library + limited set of applications
  - Very poor performance: no I/O optimizations; no parallelism

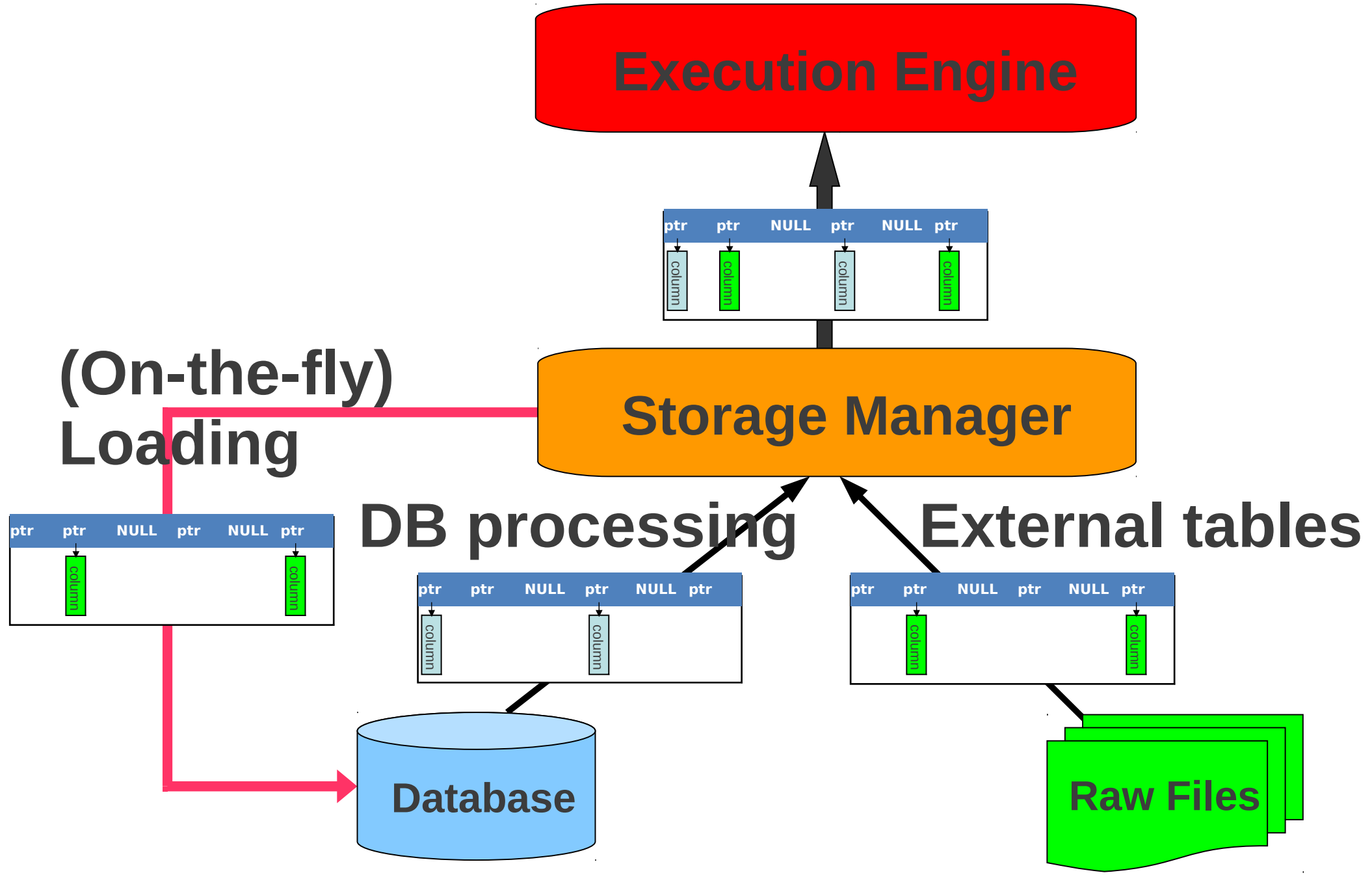
Why Not Databases?

- Data loading
  - Slow, especially for massive SAM/BAM files
  - Duplicate data (double the size)
  - Not immediate access to data
- String manipulations if no UDT support
- No SQL knowledge
- Limited support for complex aggregates, modeling, and scoring functions in SQL

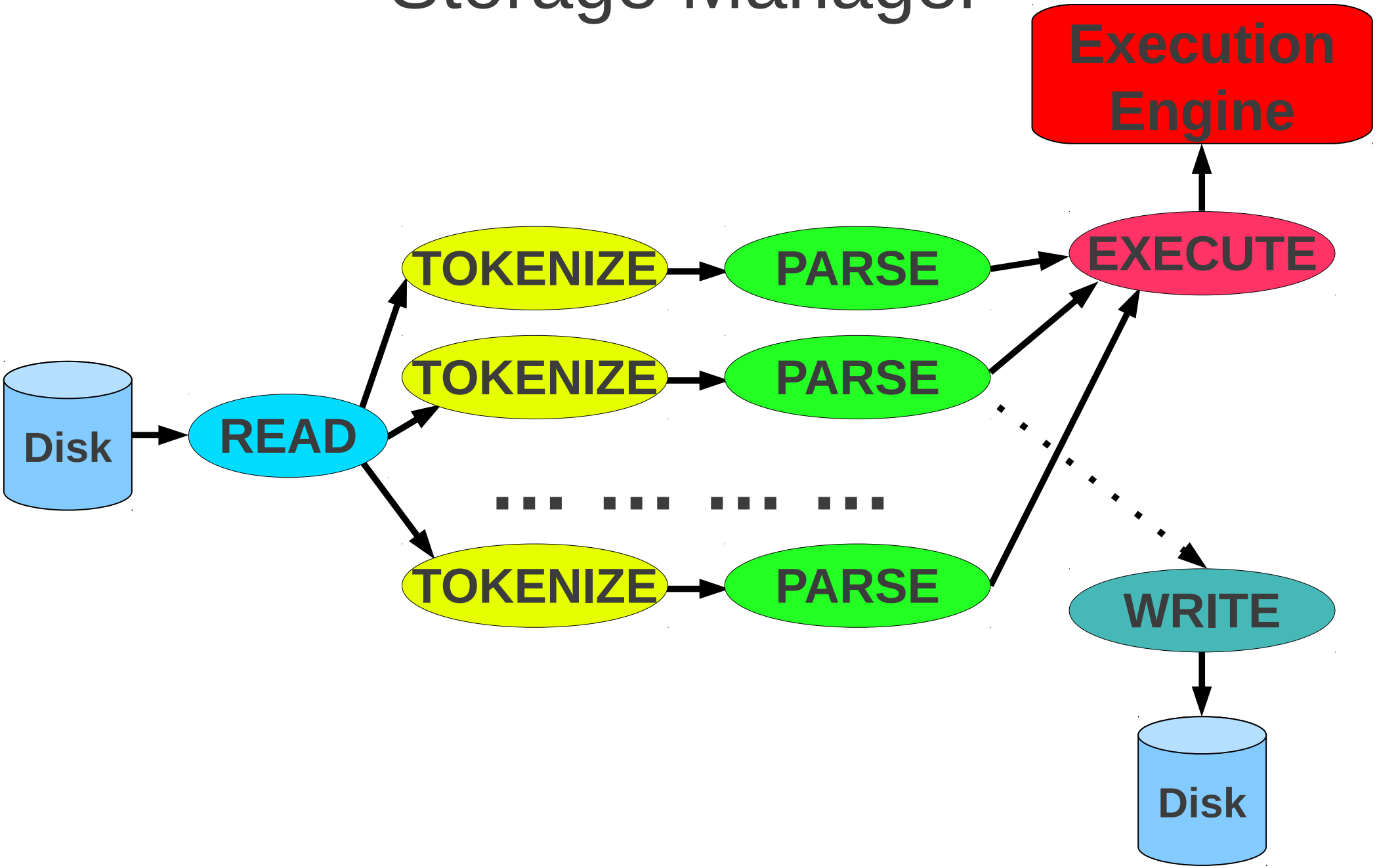
GLADE/EXTASCID



In-Situ Data Processing



Speculative Super-Scalar Pipelined Storage Manager



Experimental Results

- Setup: 16 cores; 40 GB memory; 4 X 2 TB disks RAID-0, read throughput 440 MB/sec
- Data: 1000 genome (NA12878); > 400 million reads; BAM (26 GB); SAM (145 GB)
- Task: pattern matching aggregate computation over CIGAR attribute

Task	**Execution time (sec)**	**I/O throughput (MB/sec)**
External tables (SAM)	**370**	395
External tables (BAM) + BAMtools	**2,714**	95
DB load (SAM)	945	312
DB processing	122	303
External tables (SAM) + speculative load	525	346