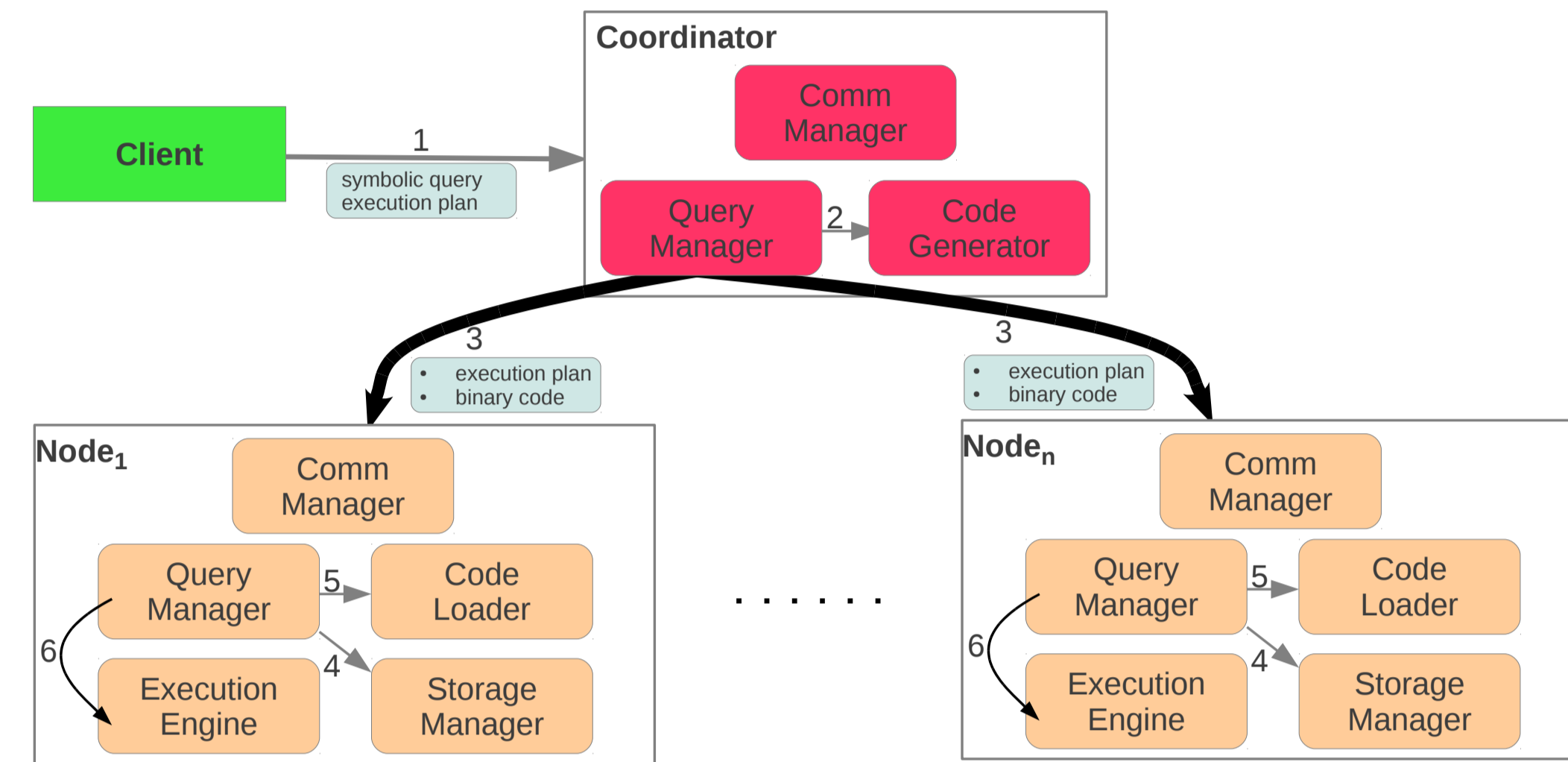


System Architecture

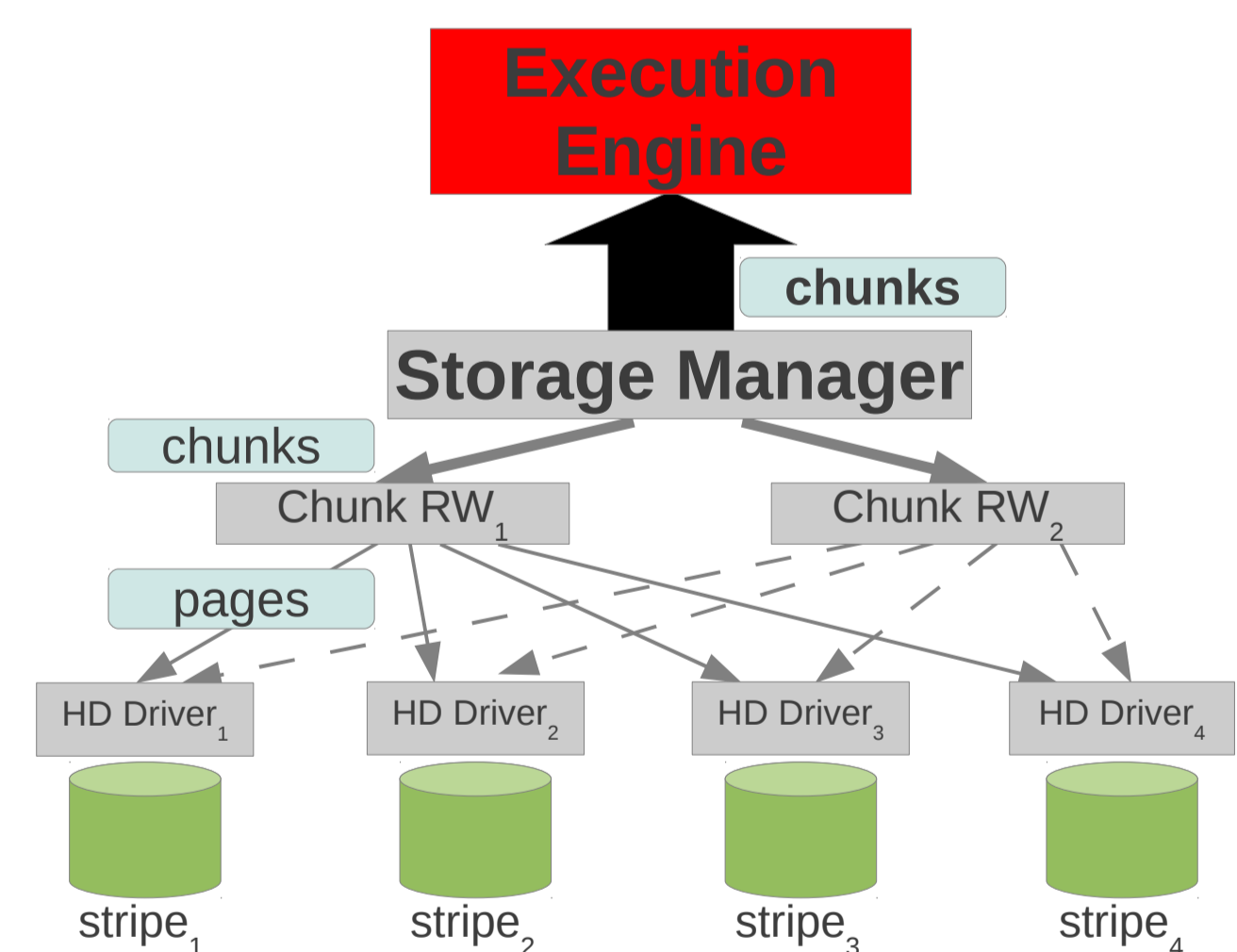
EXTASCID is a parallel system targeted at efficient analysis of large-scale scientific data



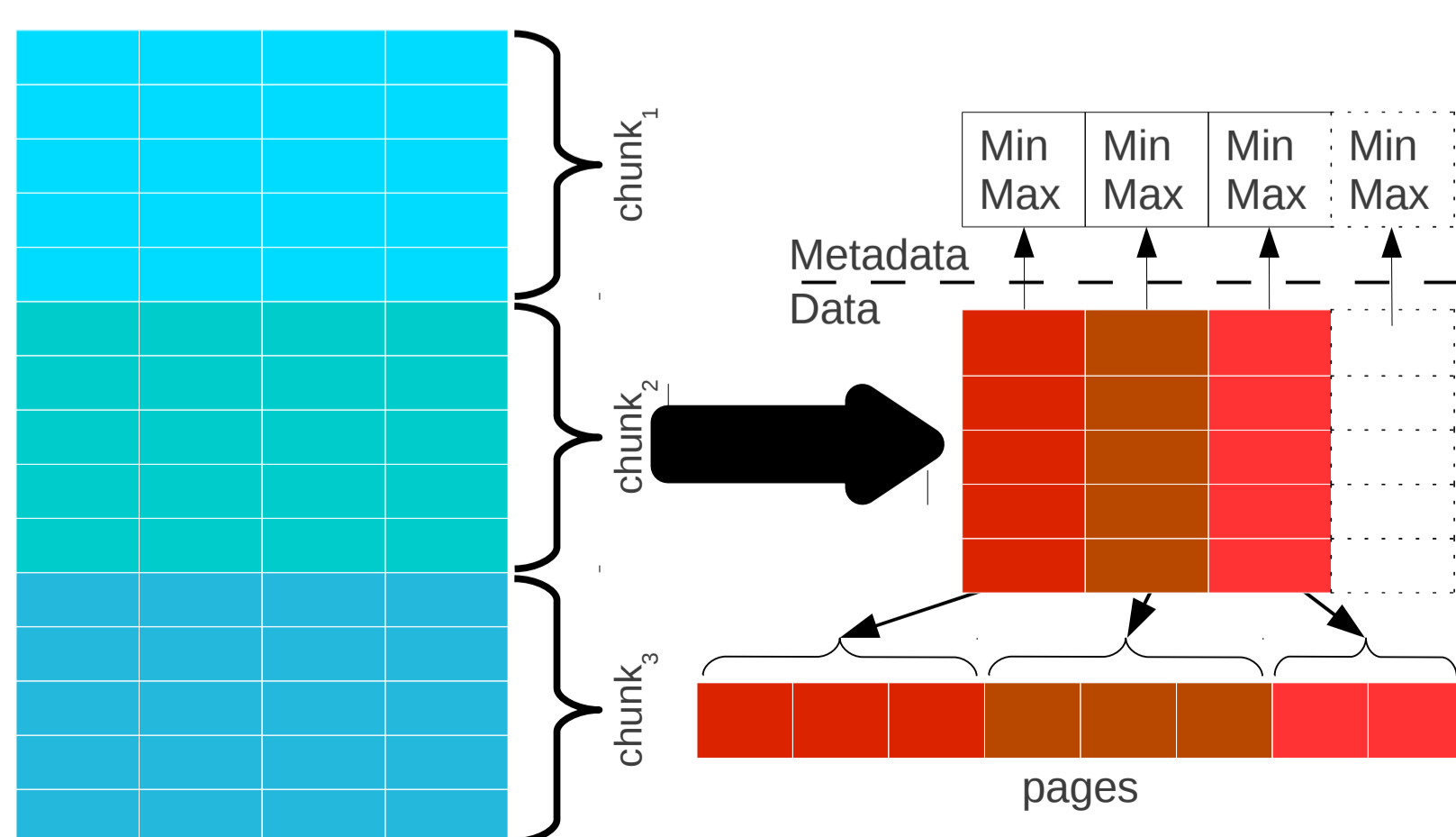
- Massive heterogeneous data: data partitioning; parallel execution; relational and array data model
- Extensible complex analytics: user code executed inside the engine; enhanced UDA interface
- Architecture independence: multi-thread (shared memory and shared disk) and inter-node (shared nothing) parallelism

Storage Manager

Push-based execution: data streaming

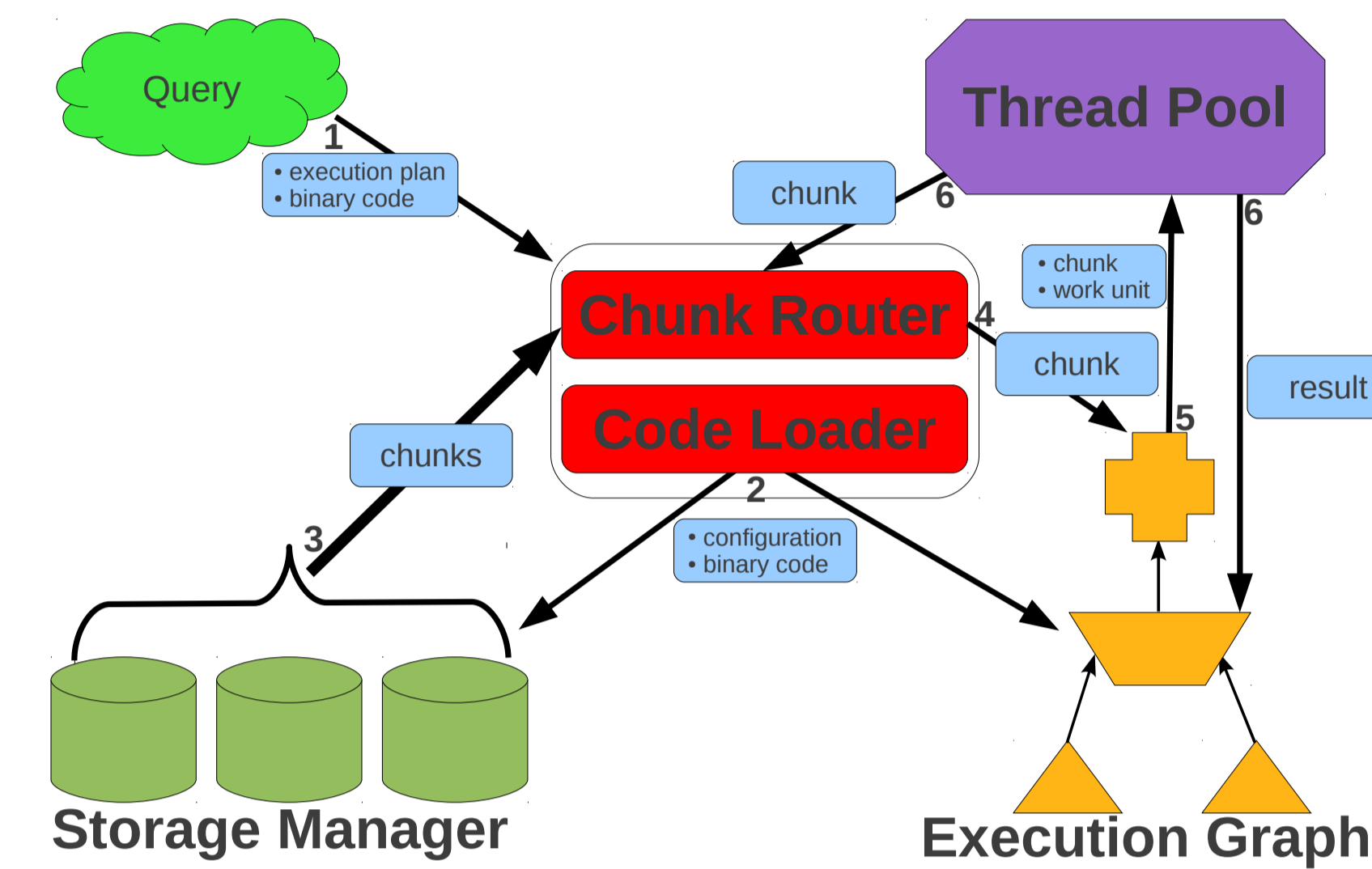


Horizontal & vertical partitioning: column-oriented chunks
 Dimension (index) suppression for dense grids

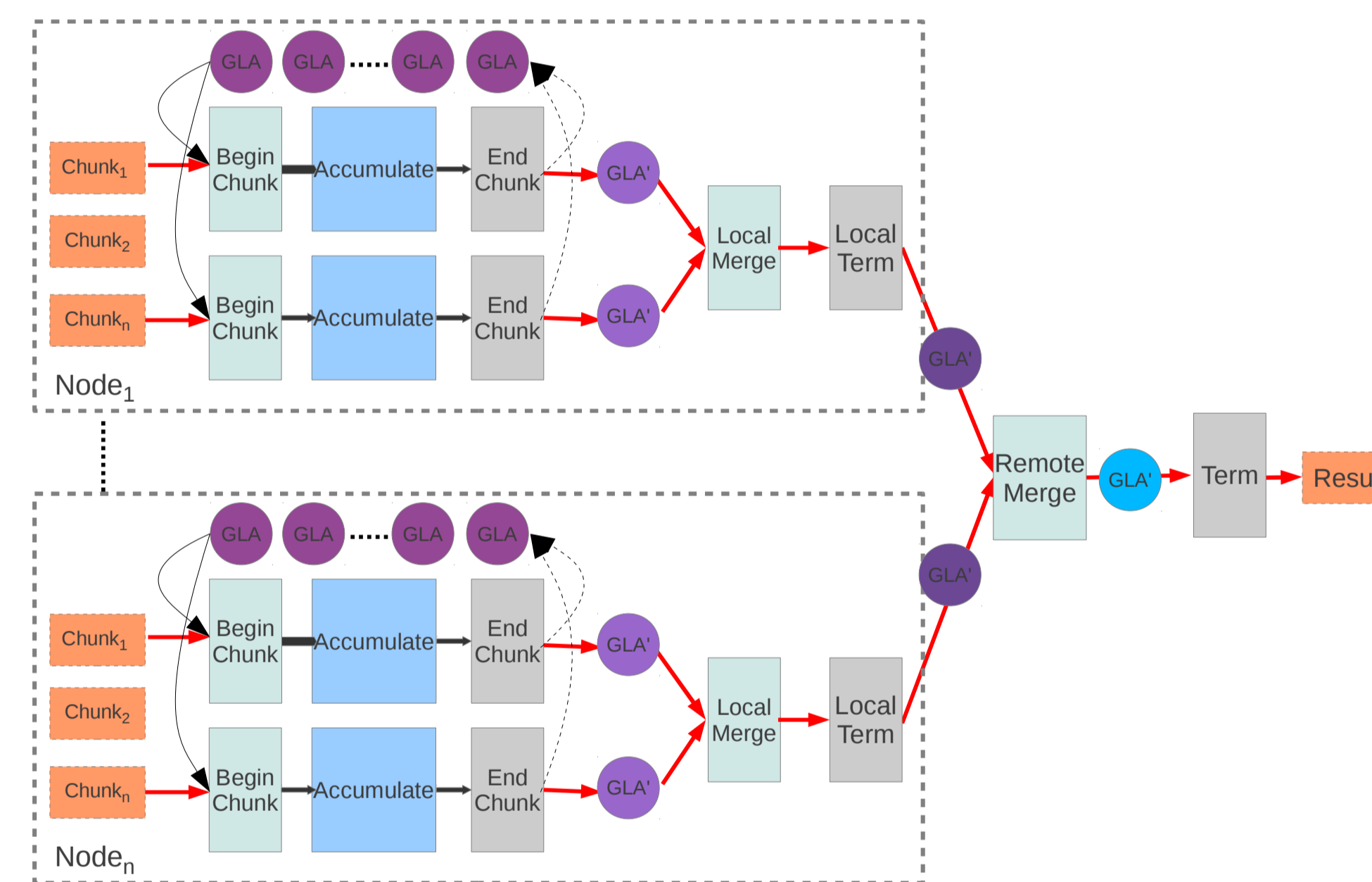


Execution Engine

Merge-oriented parallel processing

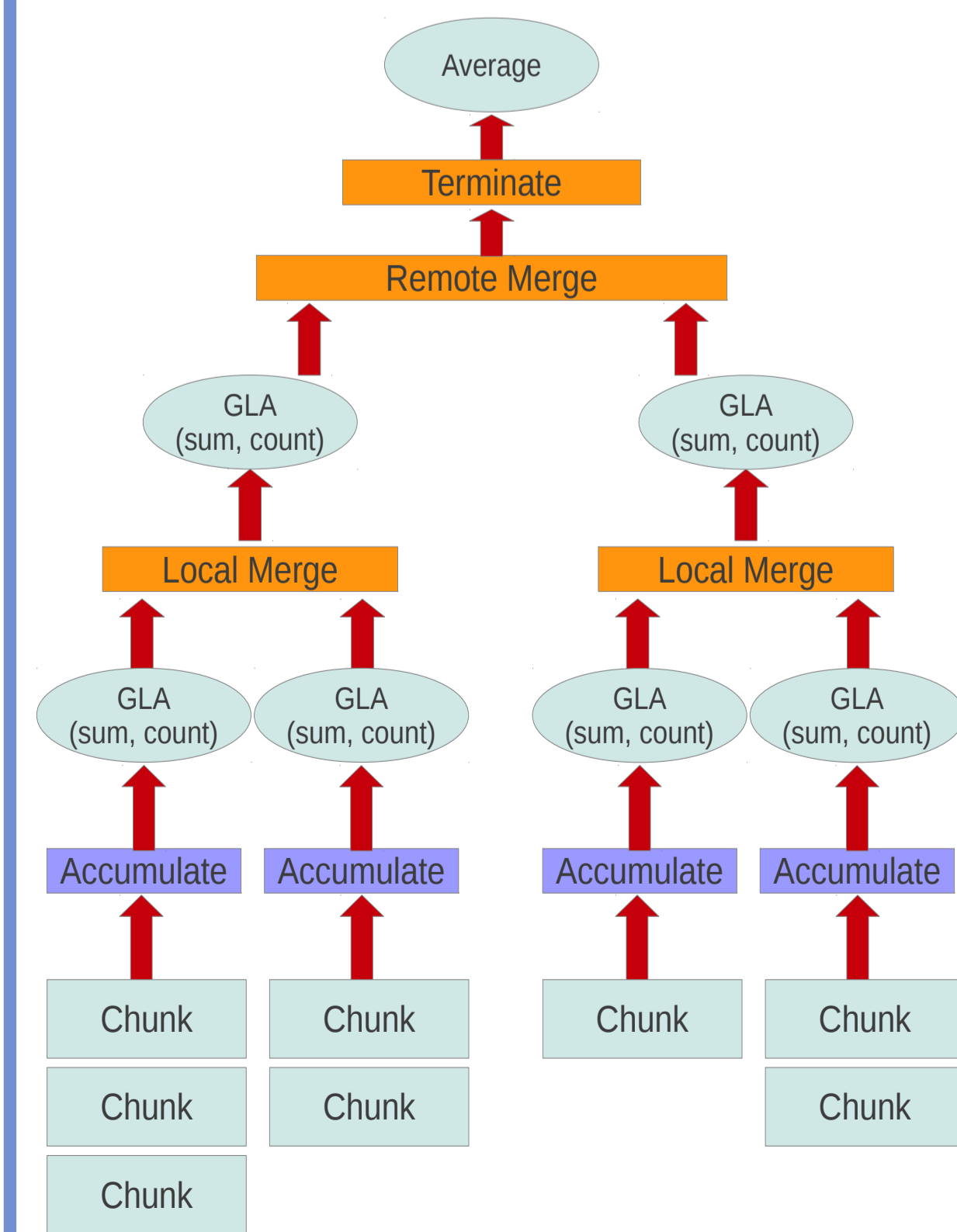


Extended User-Defined Aggregate (UDA) interface

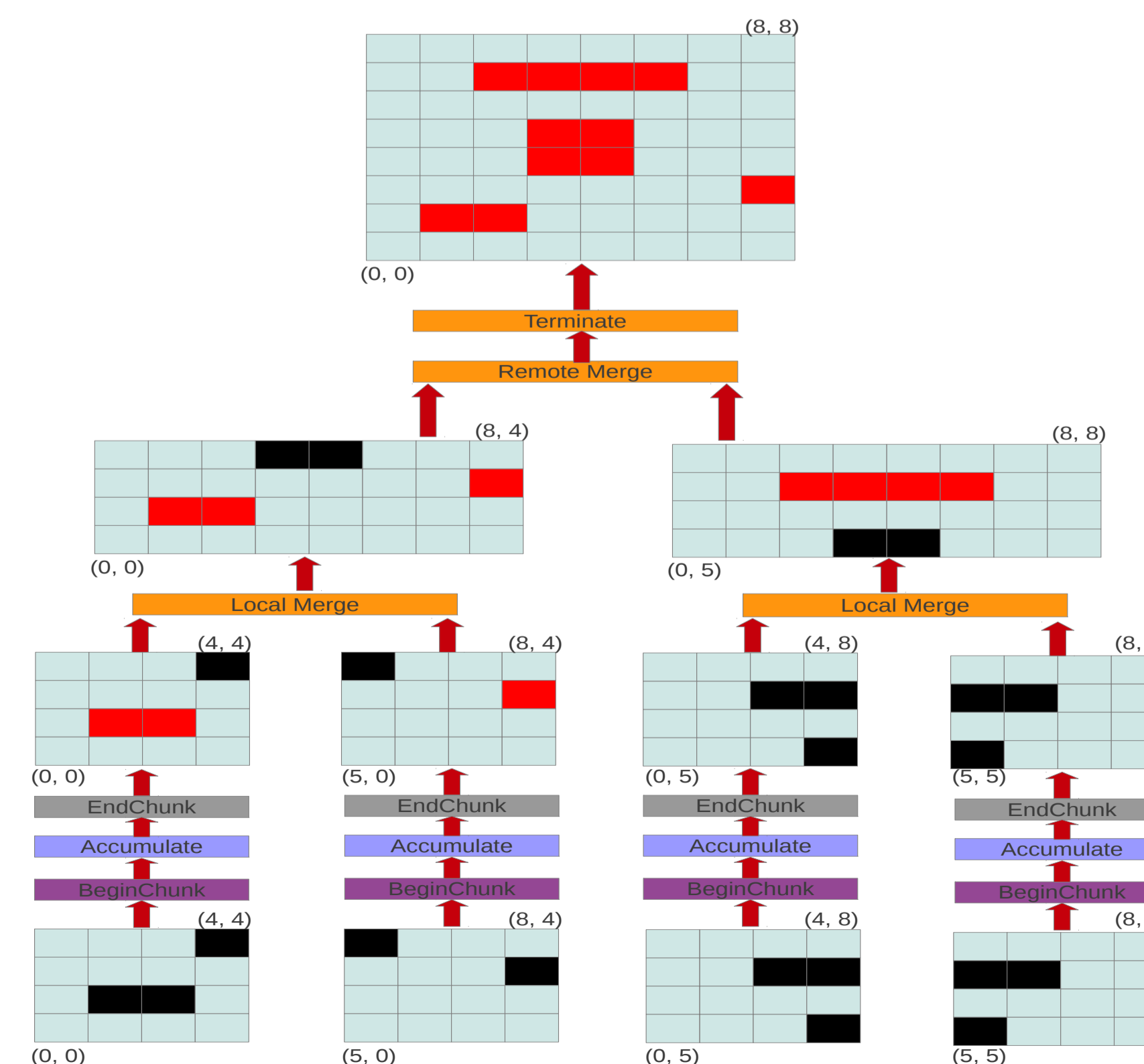


Processing Examples

Average computation



Clustering

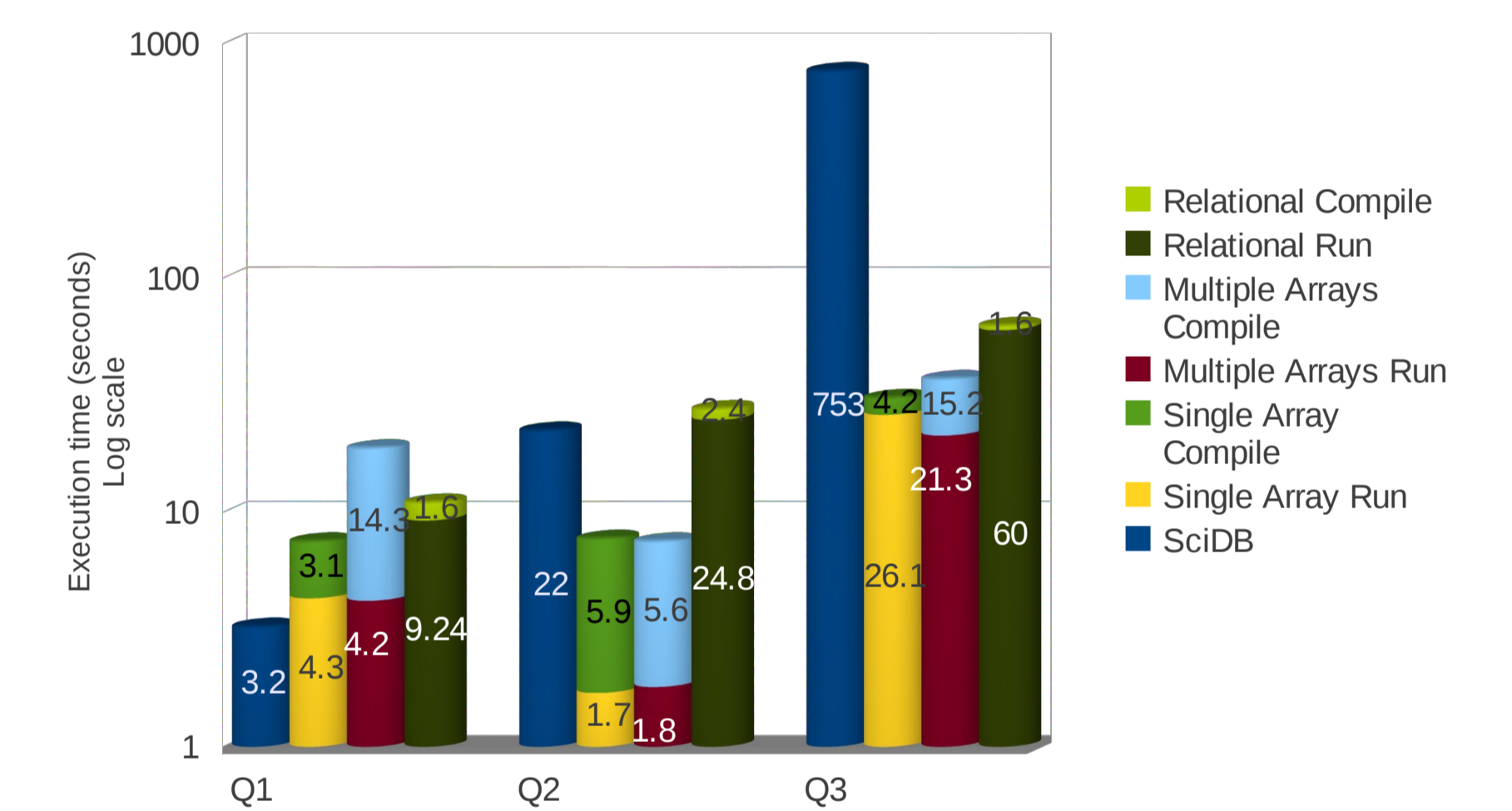


SS-DB Benchmark Experiments

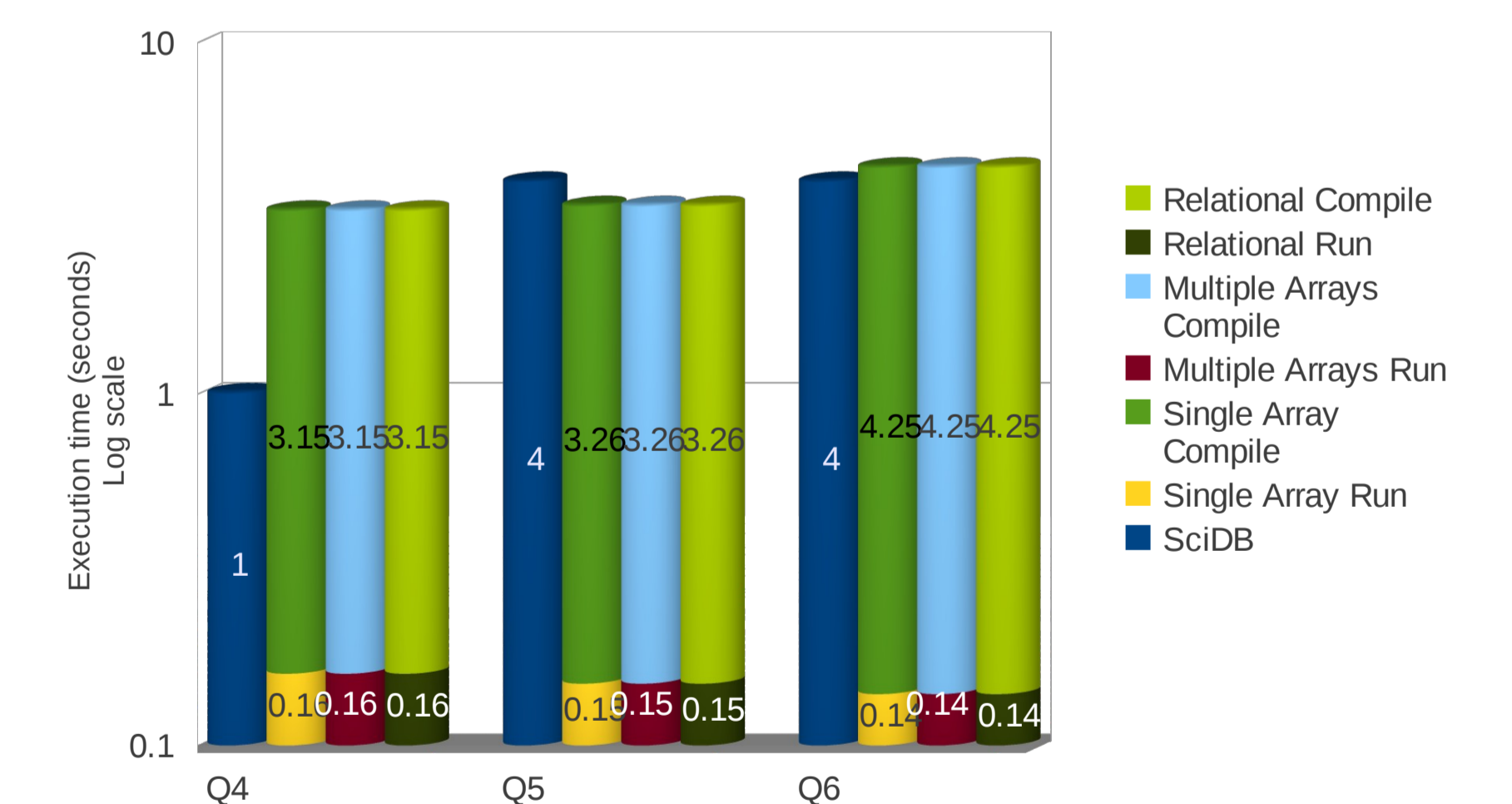
System: 9-node cluster; 1 coordinator; 8 workers
 Data loading and derived data computation

| System | Stage | Execution time [seconds] | | |
|---------------------|---------|--------------------------|-------|-------|
| | | Load | Cook | Group |
| EXTASCID Array | Compile | - | 240.5 | 5.5 |
| | Execute | 1,823 | 82.2 | 10.5 |
| | Total | 1,823 | 322.6 | 16 |
| EXTASCID Relational | Compile | - | 45.2 | 5.5 |
| | Execute | 1,829 | 797.6 | 10.5 |
| | Total | 1,829 | 842.8 | 16 |
| SciDB | - | 30,361 | 1,086 | 69 |

Queries on raw data



Queries on observations



Queries on groups of observations

