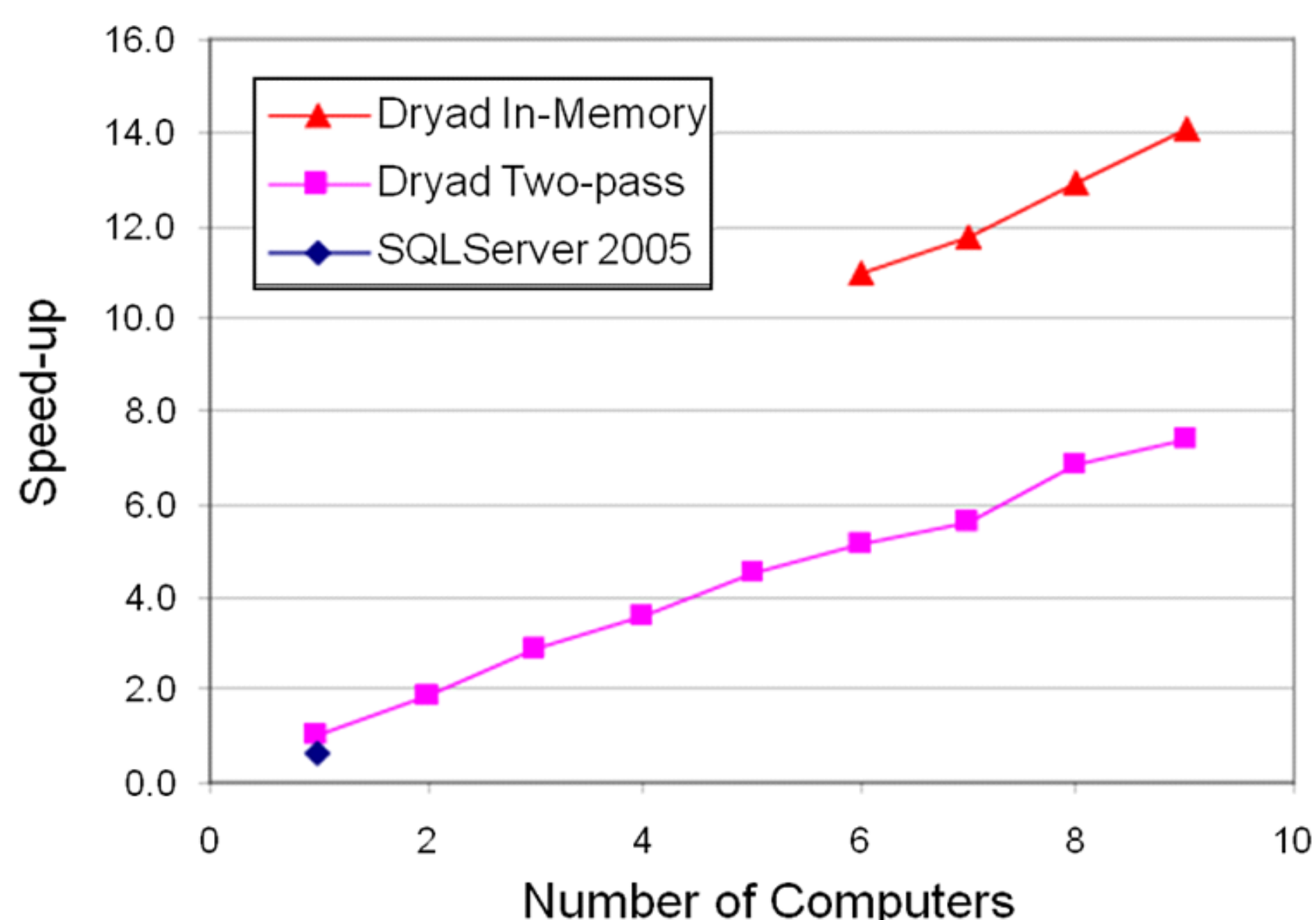
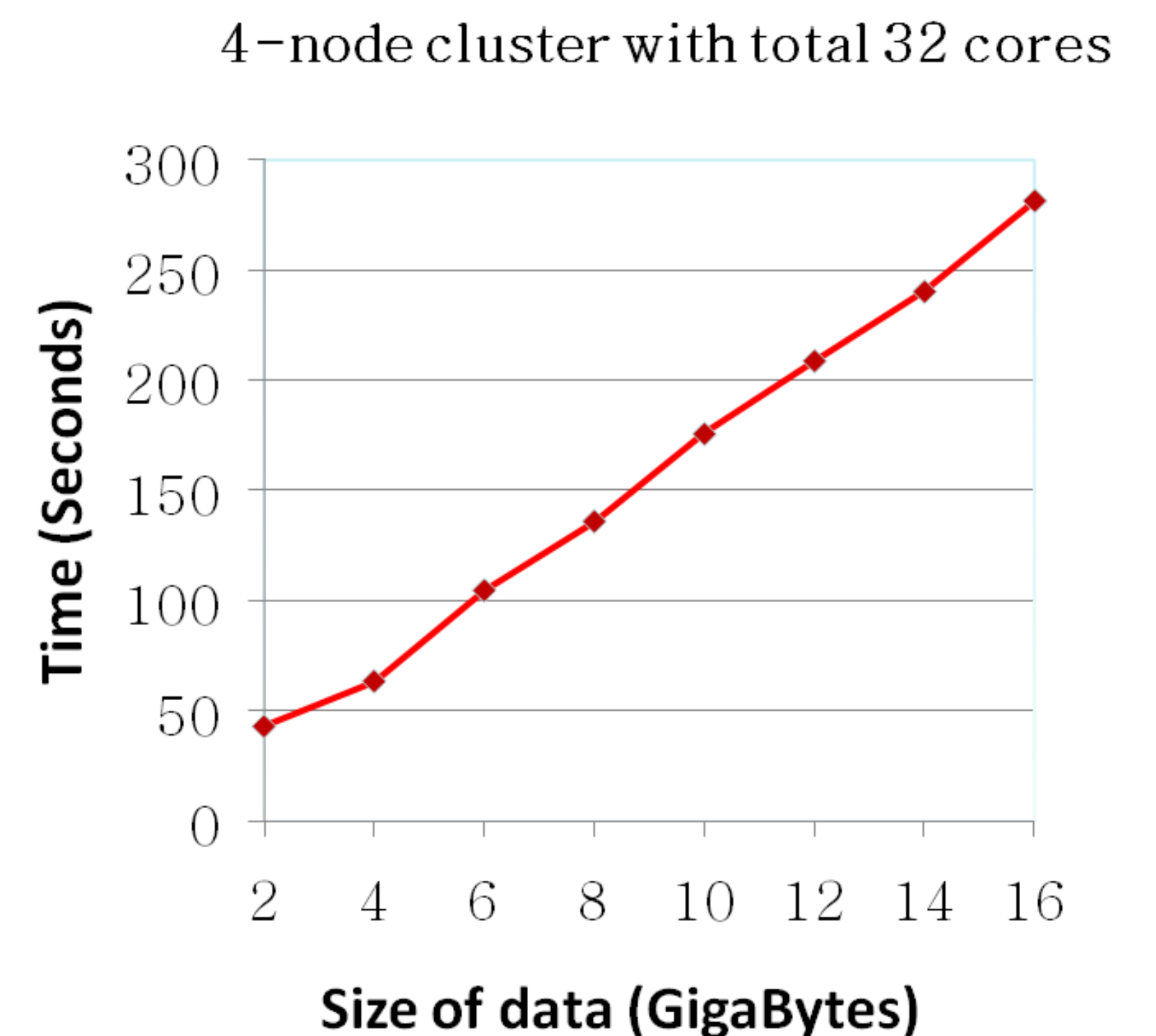


Sorting, filtering, preprocessing

- Sorting and filtering is a common data manipulation task
- Terasort, well known benchmark, time to sort time 1 TB data [J. Gray 1985]
 - In recent years, map-reduce software has set Terasort records
- DryadLINQ provides simple but powerful programming model
- Only few lines of code needed to implement Terasort

```
DryadDataContext ddc = new DryadDataContext(fileDir);
DryadTable<TeraRecord> records =
  ddc.GetPartitionedTable<TeraRecord>(file);
var q = records.OrderBy(x => x);
q.ToDryadPartitionedTable(output);
```



Sloan Digital Sky Survey

- SDSS is most ambitious astronomical survey undertaken to date, aiming to map one-quarter of the sky in detail
- SkyServer [A. Szalay, J. Gray, et al. 2001] provides internet access to SDSS data for both astronomers and science education
- **Q18**: find all objects within 30 arcseconds of one another that have very similar colors: that is where the color ratios $u-g$, $g-r$, $r-i$ are less than 0.05m
- The most time-consuming query against the SkyServer database, about 150 million comparisons

Probabilistic Index Maps (PIM)

[Jojic and Caspi CVPR 2004]

- Unsupervised extraction of natural parts from a collection of related signals, using PIM
- Generate sets of common features from images

