Parallel and Distributed Computing in R

Kanak Choudhury

MapReduce Technique

- MapReduce originally referred to the proprietary Google technology
- A popular open-source implementation is Apache Hadoop
- A MapReduce program is composed of a Map() procedure that performs on each part of data and a Reduce() method that performs a summary operation

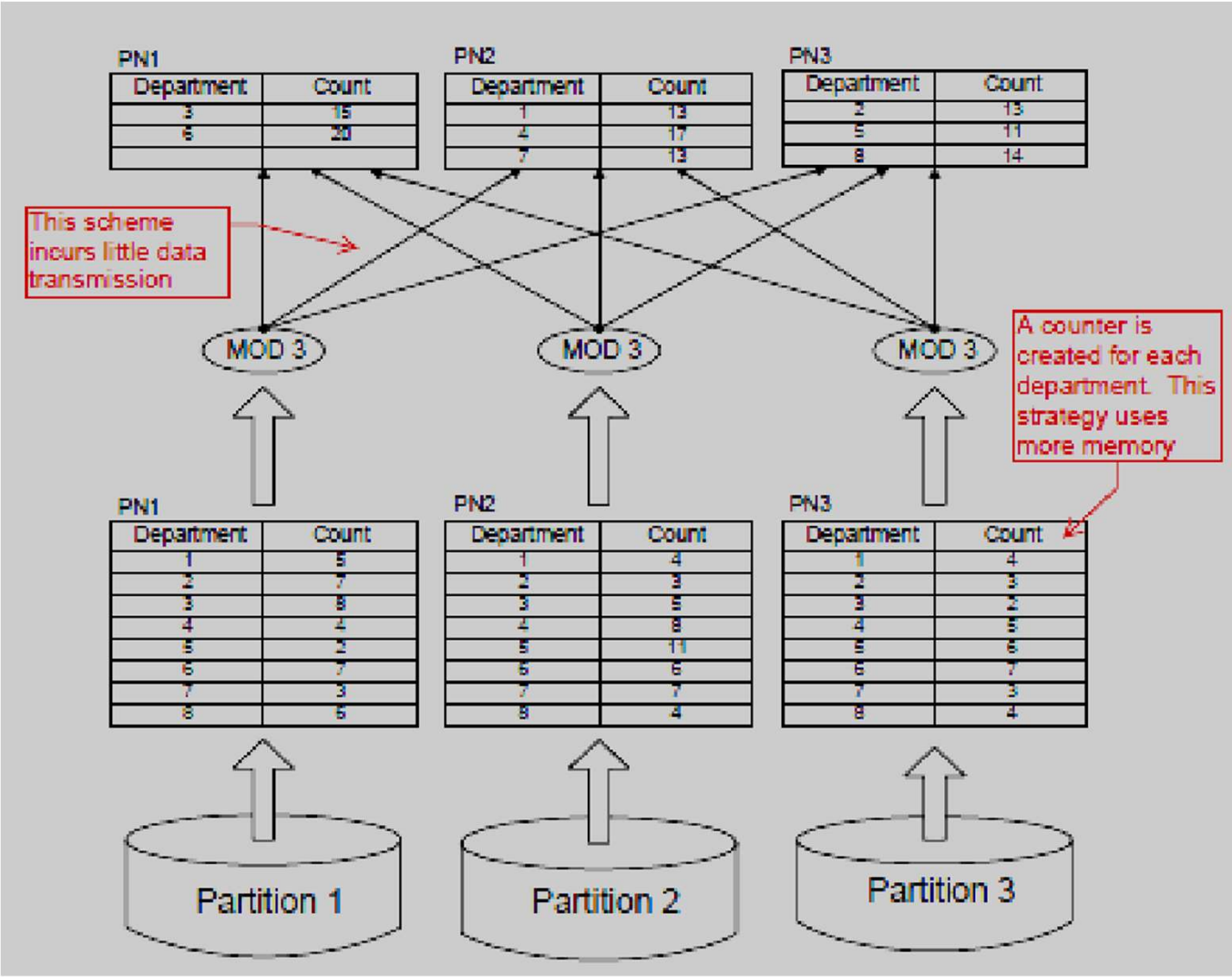
MapReduce Technique

- **"Map" step:** "map()" function applies to each part of data by each worker node.

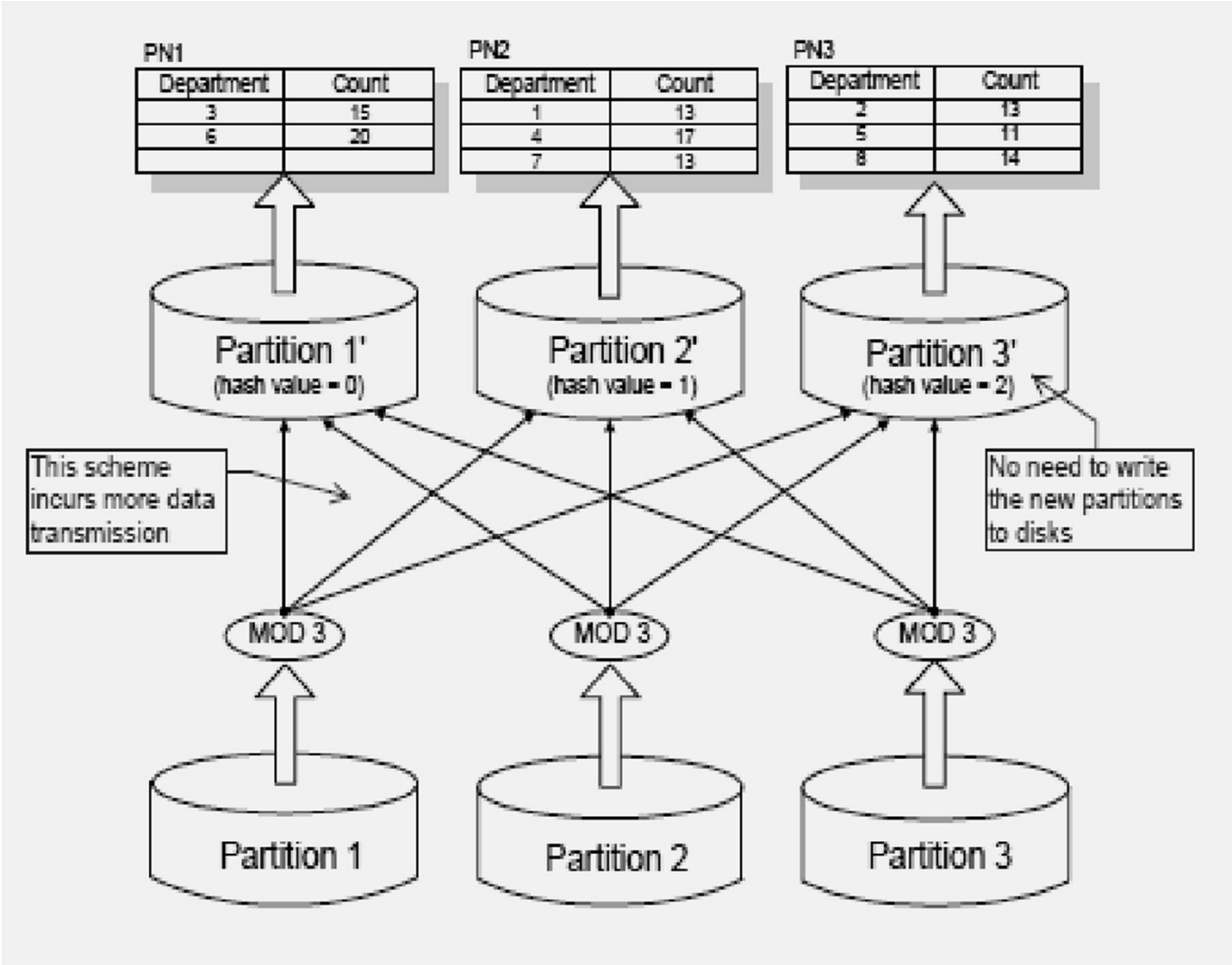
- **"Shuffle" step:** Output obtained by “map()” function redistribute such that all data with same key is located on the same worker node.

- **"Reduce" step:** Worker nodes now process each group of output data, per key, in parallel.

Example: MapReduce

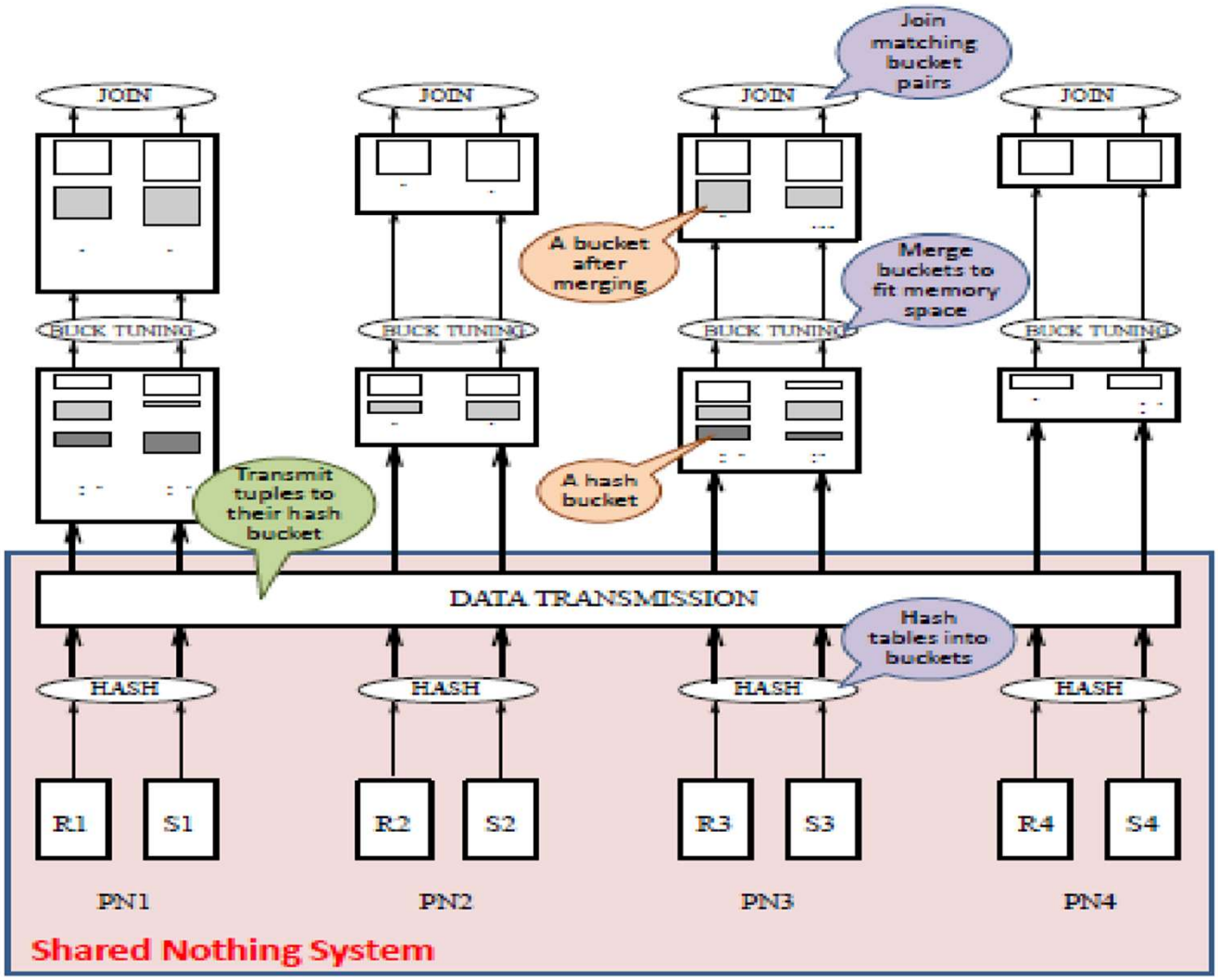


Example: MapReduce

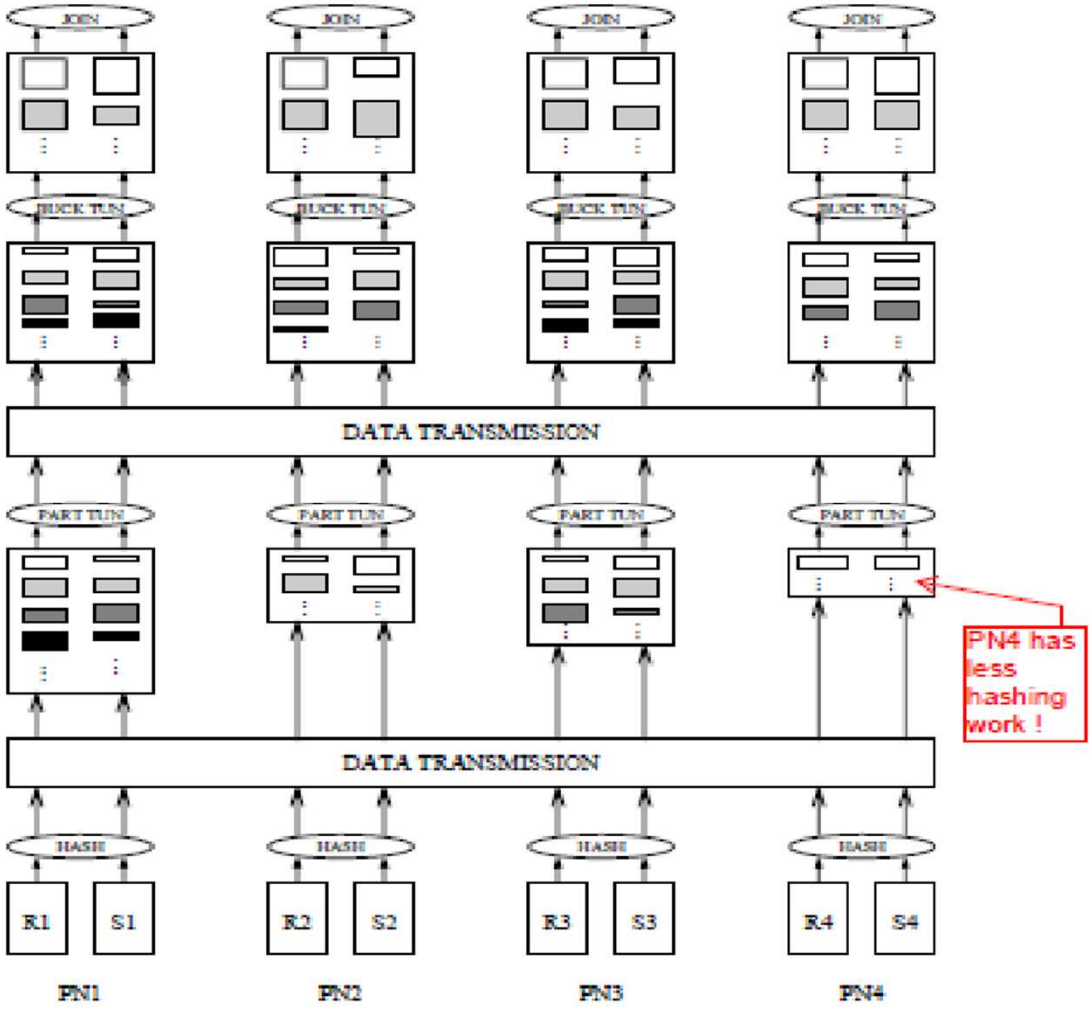


Parallel Computing

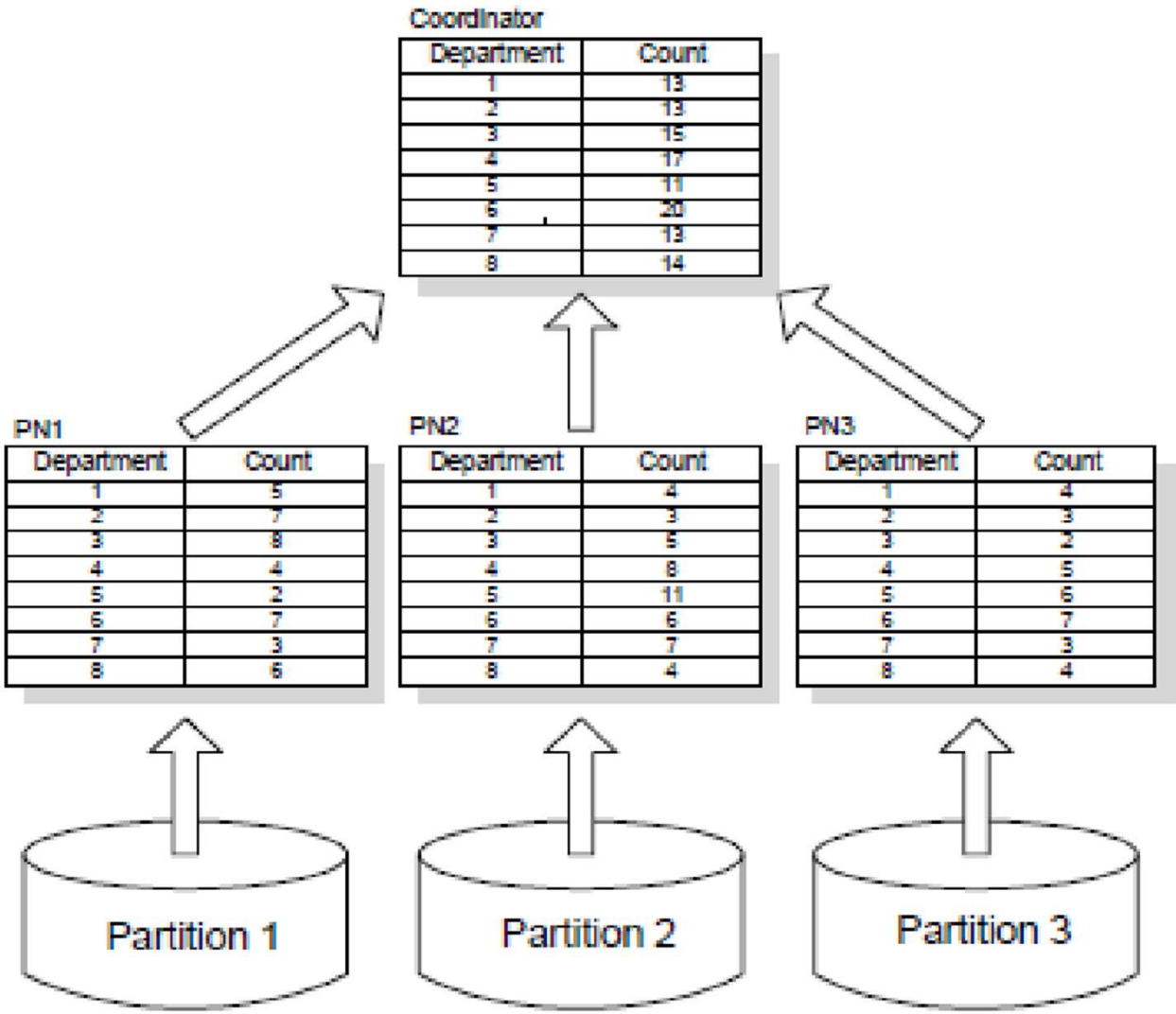
Divide large problems into smaller ones, solve independently (many at the same time) and results are combined afterwards



Parallel Computing (Load Balancing)



Example: Parallel Computing



Parallel Computing in R

- https://cran.r-project.org/web/views/HighPerformanceComputing.html

R Packages

- Parallel (SOCK (Socket) concept)
- doParallel
- Snow
- doRedis (MPI (message passing interface) concept) (need to setup redis server)

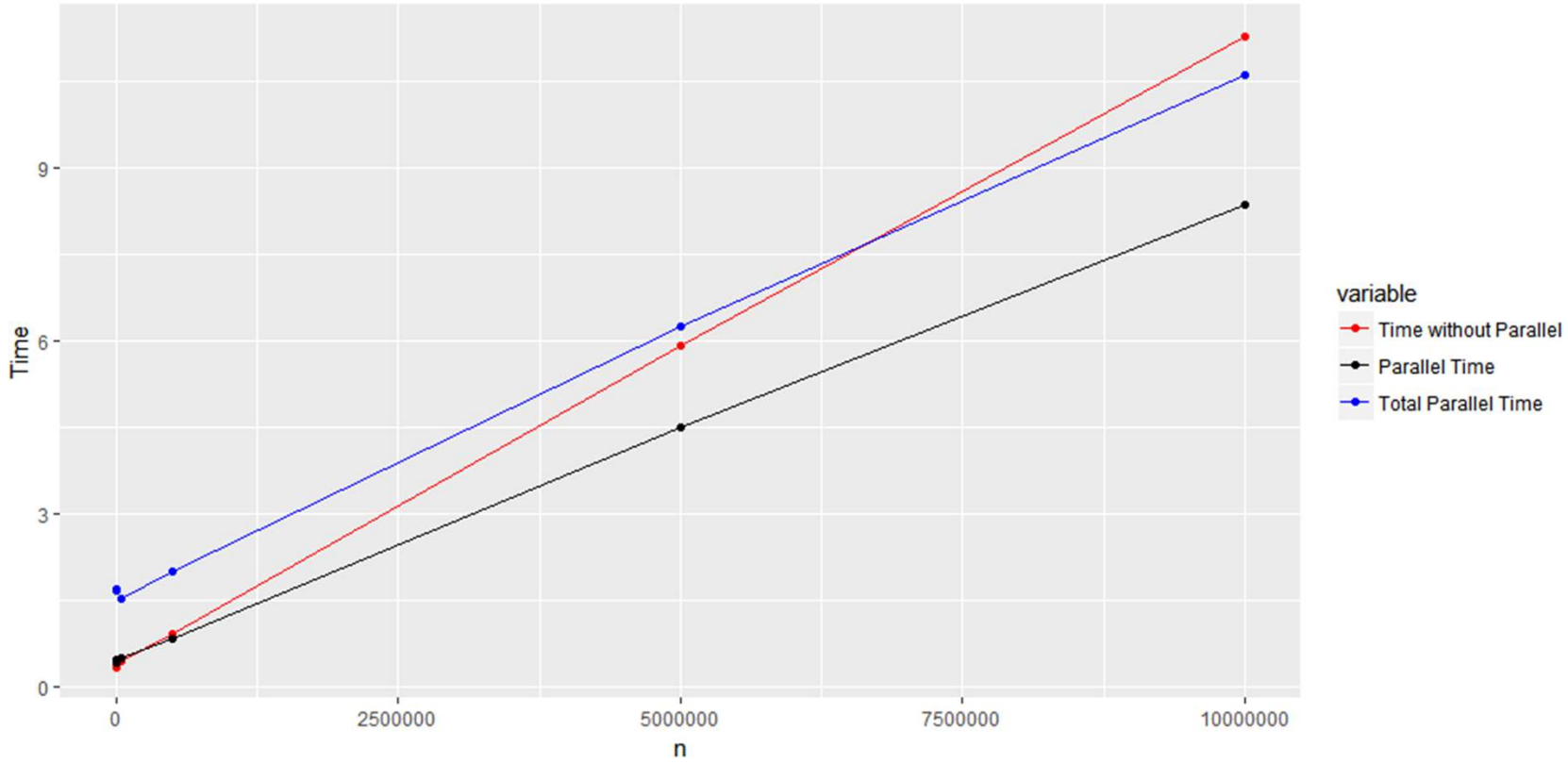
  https://github.com/MicrosoftArchive/redis/releases
- Rmpi
- doMPI
- plyr (it uses the parallelization concept)

Parallel Functions

- clusterCall
- clusterApply
- clusterApplyLB
- clusterEvalQ
- clusterExport
- clusterMap
- clusterSplit

- parLapply
- parSapply
- parApply
- parRapply
- parCapply
- parLapplyLB
- parSapplyLB

Parallel Computing Time



R Package for GPU

- gpuR
- cudaBayesreg
- Rgpu (speed up bioinformatics analysis by using GPU) https://trac.nbic.nl/index/
- RPUD
- RPUDPLUS
- RPUSVM
- Accelerate R Applications with CUDA (Compute Unified Device Architecture) https://devblogs.nvidia.com/accelerate-r-applications-cuda/

Parallel Computing: Hadoop

- RHIPE (http://deltarho.org/docs-RHIPE/#the-r-session-server-and-rstudio)
- https://www.dezyre.com/article/r-hadoop-a-perfect-match-for-big-data/292
- **rMR**
- **RProtoBuf**

Machine Learning Packages

- caret
- h2o (computes parallel distributed machine learning algorithms)
  GLM, GBM, RF and NN within various cluster environments
- parallelSVM
- ParallelForest
- e1071

Reference

1. https://en.wikipedia.org/wiki/MapReduce
2. http://www.cs.ucf.edu/~kienhua/classes/
3. https://cran.r-project.org/web/views/HighPerformanceComputing.html
4. https://github.com/MicrosoftArchive/redis/releases
5. https://trac.nbic.nl/index/
6. http://deltarho.org/docs-RHIPE/#the-r-session-server-and-rstudio
7. https://www.dezyre.com/article/r-hadoop-a-perfect-match-for-big-data/292

Thank you!