DSIMBench: A benchmark for microarray data using R

Shicai Wang

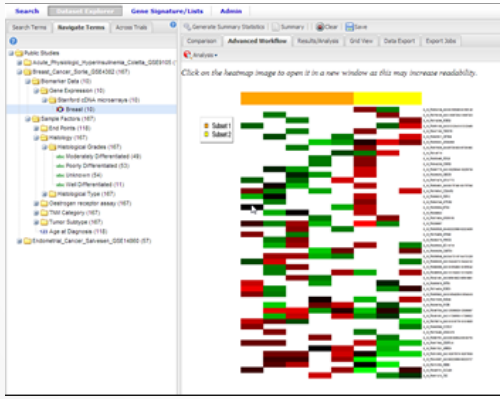
Imperial College London

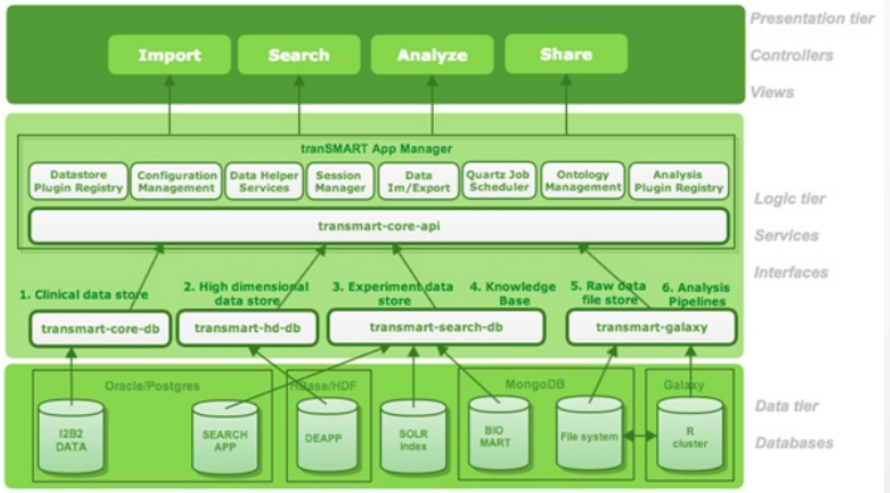
Outline

- Motivation
- Benchmark design
- Experimental evaluation

Motivation

- tranSMART
  - Translational medical knowledge management
  - Collaborative projects
  - > 100,000,000 records of gene expression and SNP tables require Big Data solution





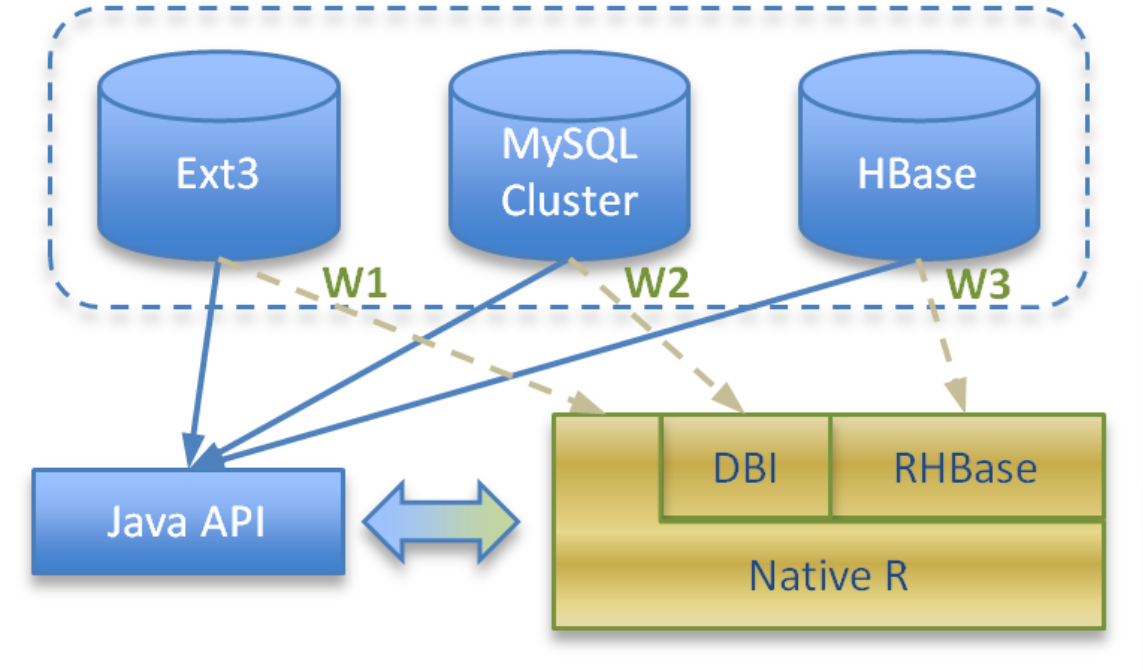
Design

- Workflows

Workflows	Data loading	Computation	Data source	Parallel method
W1	Single process	N/A	ext3	N/A
W2	Single process	N/A	HBase	N/A
W3	Single process	N/A	MySQL Cluster	N/A
W4	N/A	Single process	N/A	Native R
W5	N/A	Multiple cores	N/A	MPI
W6	N/A	Multiple cores	N/A	MapReduce
W7	Multiple processes	Multiple cores	Best DB	MPI
W8	Multiple processes	Multiple cores	RHBase	MapReduce

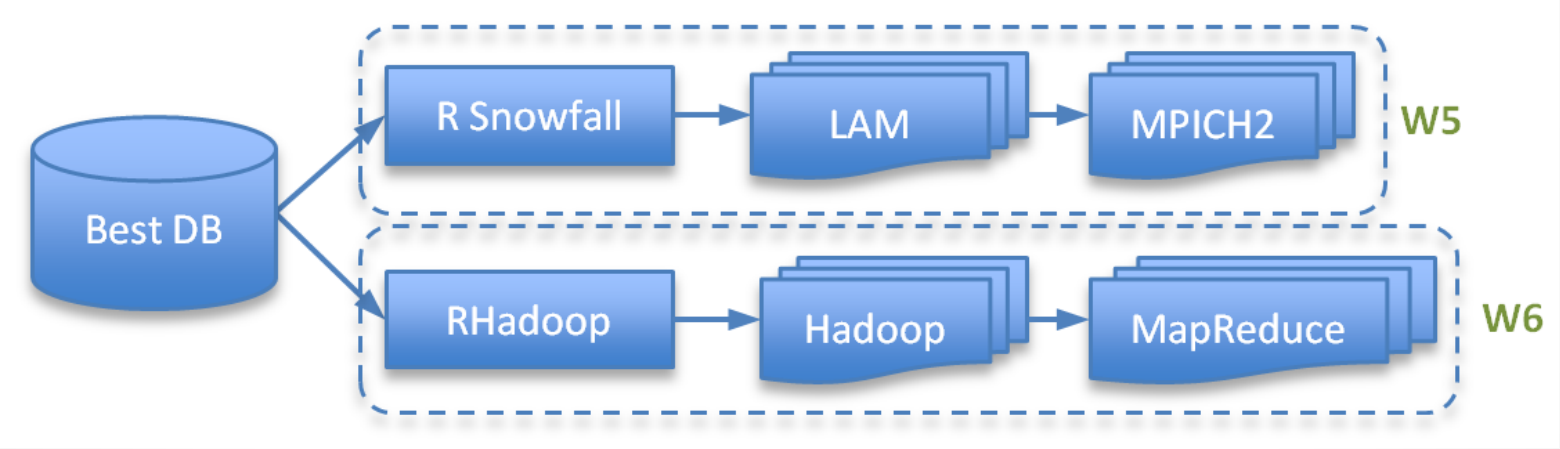
Design

- Data distribution tests (W1-W3)
  - Java API to evaluate the affect from the R middleware.



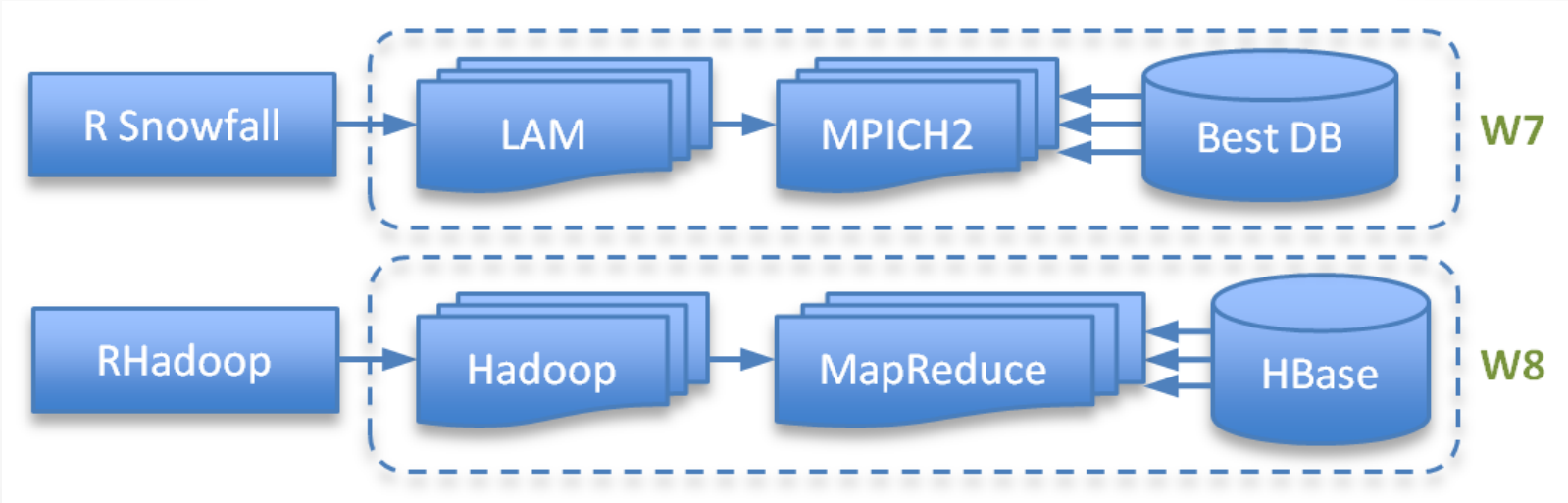
Design

- Computation tests (W4-W6)
  - W4 (native R): baseline.



Design

- Mixed tests (W7, W8)



Applications

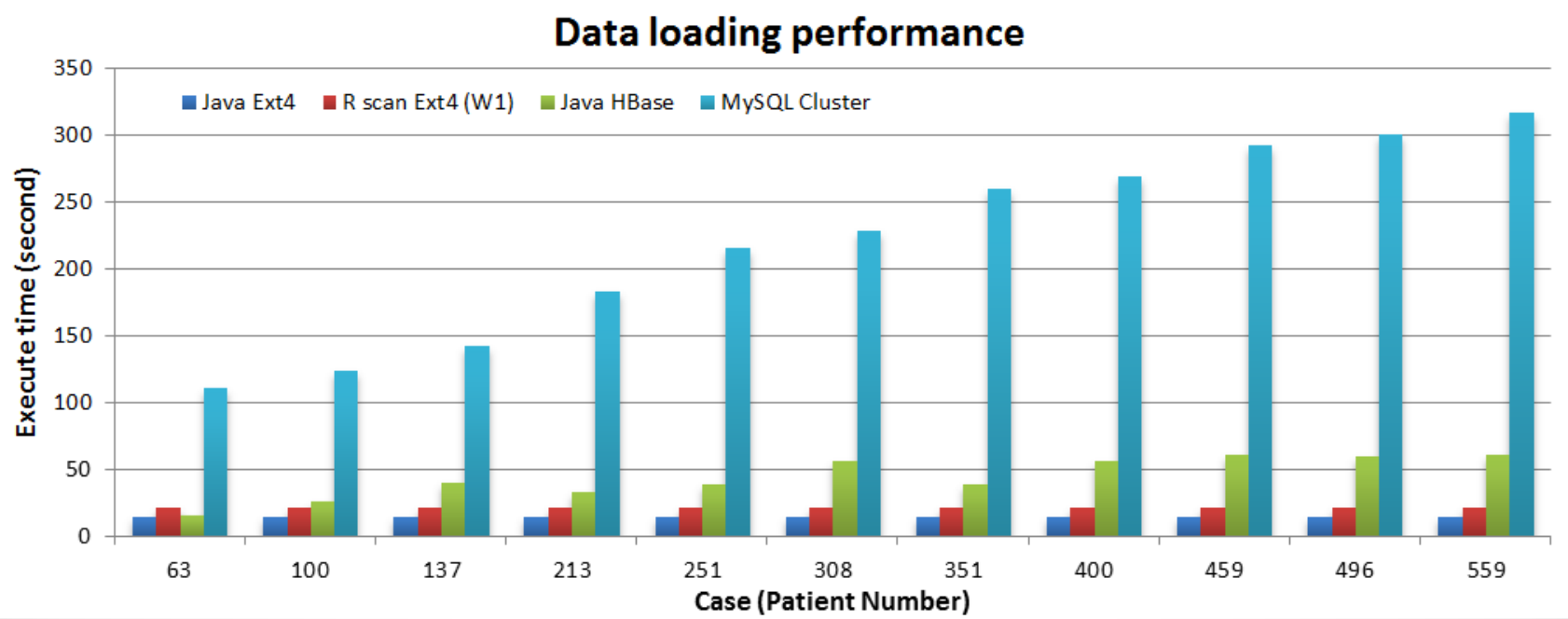
- Data loading test
  - Marker selection
- Computation test
  - Hierarchical Clustering
- Mixed test
  - Marker selection & Hierarchical Clustering

Experimental Evaluation

- Enviroment (VM in IC Cloud platform)
  - Vanilla R: one VM (32 CPU, 32 GB memory)
  - LAM/MPI: four VM (8 CPU, 8 GB memory)
  - Hadoop: four VM (8 CPU, 8 GB memory)
- Datasets
  - NCBI GSE24080 (559 subjects, 54675 probeset)
  - NCBI GSE2109 (2096 subjects, 54675 probeset)

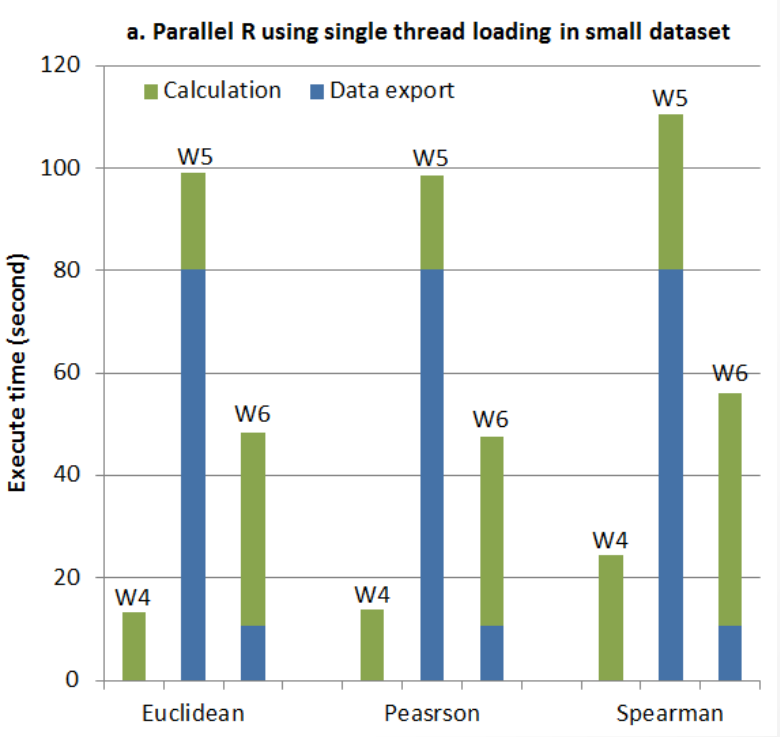
Experimental Evaluation

- Data loading tests



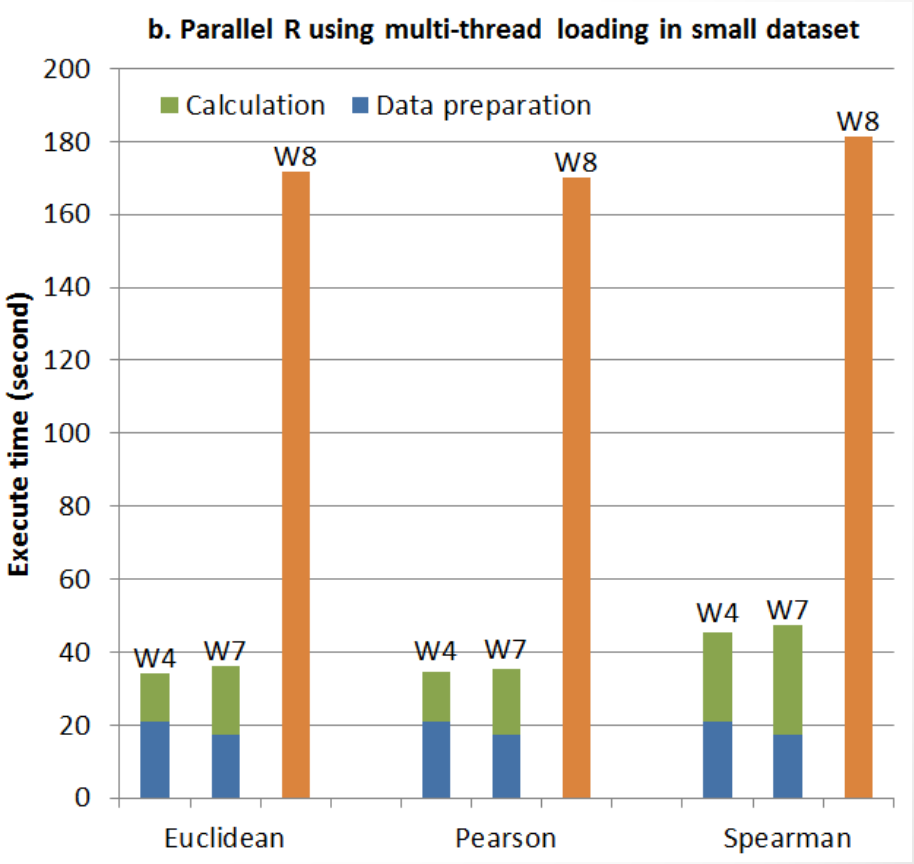
Experimental Evaluation

- Computation tests (Smaller dataset)
  - Native R performs best
  - Large data transfer overhead in Snowfall in W5 (blue)
  - Pure computation time (green) MPI < MapReduce



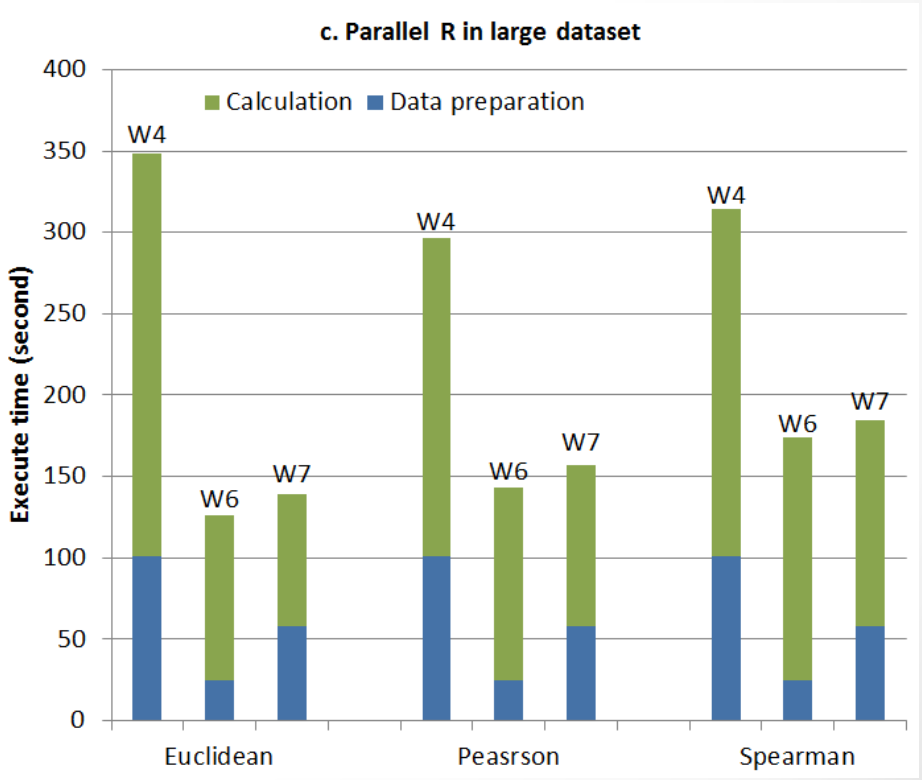
Experimental Evaluation

- Mixed tests (smaller dataset)
  - W4 (native R) and W7 (MPI) perform much better than W8 W8 suffers from rhbase and thrift overhead



Experimental Evaluation

- Summary of the best performance in larger datasets
  - W6 performs best
  - Computation time: W6 > W7



Conclusion

- Two classic microarray computation
- Eight different parallel R workflows to evaluate the Big Data solution
  - 3 data loading tests
  - 3 computation tests
  - 2 mixed tests

Q & A

Thanks!