# Database Systems
Sort-Merge and Hash Join

Chaitanya Patel - 201501071

- Configuration of System
  - HP Pavilion p26x
  - Intel i5 5th Gen 2.2 GHz
  - 8 GB RAM
  - 1 TB hard drive

- Input Relation R and S
  - R(X, Y) and S(Y, Z) is randomly generated
  - Both R and S have 10,000 rows (approx 1 MB each)
  - Join attribute value is integer in range [0, 500)
  - So there will be approx 200 rows in each relation with same Y attribute value
  - Thus the output natural join relation will have approx 20,000,000 rows (approx 300 MB)
  - According to the assignment statement, each block is assumed to carry maximum 100 rows

- Sort-Merge-Join
  - $B(R) + B(S) <= M*M$
  - For our input file, $M >= 46$
  - With respect to change in M, Since number of I/Os don't change, the execution time is not changing drastically.
  - Slight increase of performance increase with increase in M can be attributed to fewer number of sublists to process. Due to larger sublists, the I/Os are happening sequentially accessing sequential data on disk, hence slight performance increase.

M : Number of Main Memory Blocks	Execution Time
46	real 0m14.089s user 0m9.808s sys 0m0.544s
47	real 0m13.837s user 0m9.056s sys 0m0.444s
50	real 0m13.985s user 0m10.184s sys 0m0.468s
60	real 0m13.581s user 0m9.376s sys 0m0.404s
70	real 0m12.450s user 0m9.400s sys 0m0.436s
80	real 0m11.958s user 0m8.836s sys 0m0.392s
90	real 0m11.752s user 0m8.724s sys 0m0.528s
100	real 0m11.454s user 0m9.020s sys 0m0.416s

Sort-Merge-Join
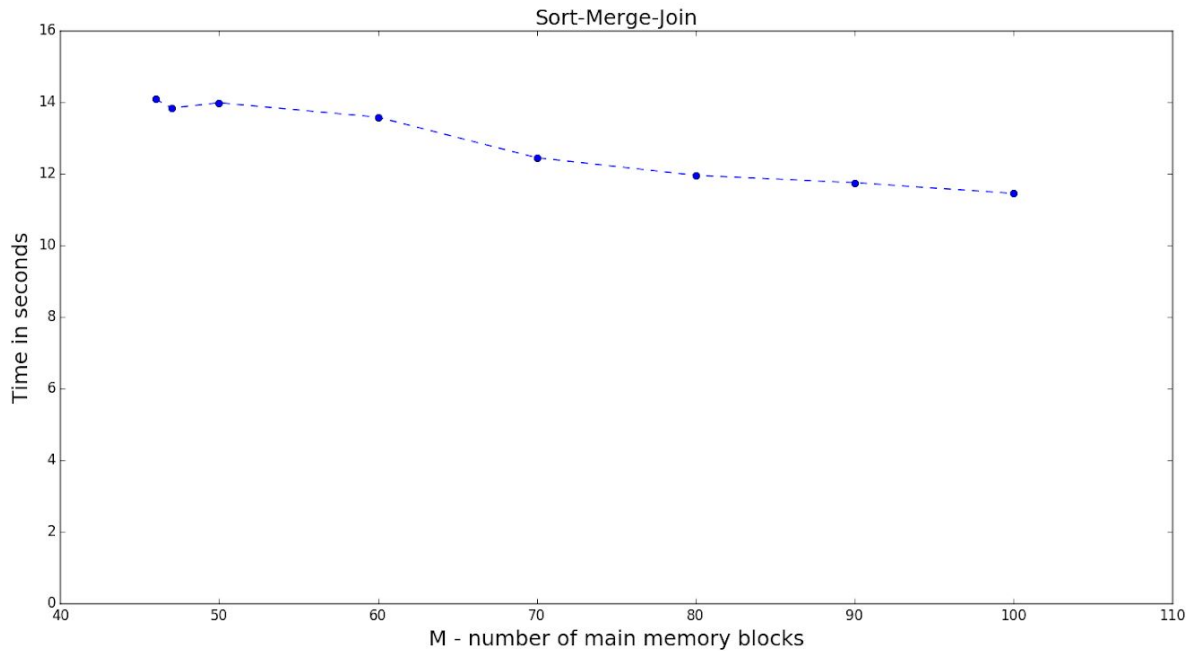
- Hash Join
  - min(B(R), B(S)) <= M*M
  - For our input file, M >= 33
  - Since the input file is randomly generated, it won't follow uniform distribution of tuples over all buckets. In our input file, minimum needed M comes out to be 40 i.e. M >= 40
  - With respect to change in M, Since number of I/Os don't change, the execution time is not changing drastically.
  - On finer level, performance is first slightly increasing with respect to increase in M because of fewer number of buckets. With more high M value, performance doesn't increase because increased number of buckets cause more infrequent disk I/Os.

M : Number of Main Memory Blocks	Execution Time
33	real 0m14.422s user 0m10.992s sys 0m0.404s
40	real 0m13.595s user 0m10.272s sys 0m0.400s

50	real 0m13.531s user 0m10.164s sys 0m0.460s
60	real 0m13.469s user 0m10.640s sys 0m0.400s
70	real 0m12.921s user 0m10.036s sys 0m0.424s
80	real 0m12.077s user 0m10.152s sys 0m0.516s
90	real 0m12.456s user 0m10.024s sys 0m0.452s
100	real 0m12.669s user 0m10.044s sys 0m0.392s

Hash-Join