Ferhat Elmas 214805

Advanced Databases Hw-2

1) **NLJ:**

- (Peter, 1) x (1, Advanced Databases)
- (Peter, 1) x (2, Advanced Databases)
- (Peter, 1) x (2, Quantitative Risk Assessment)
- (Peter, 1) x (4, Diversity Management)
- (Peter, 1) x (3, Microeconomics)
- (Peter, 1) x (1, Marketing and Public Relations)
- (Kate, 2) x (1, Advanced Databases)
- (Kate, 2) x (2, Advanced Databases)
- (Kate, 2) x (2, Quantitative Risk Assessment)
- (Kate, 2) x (4, Diversity Management)
- (Kate, 2) x (3, Microeconomics)
- (Kate, 2) x (1, Marketing and Public Relations)
- (Maggie, 3) x (1, Advanced Databases)
- (Maggie, 3) x (2, Advanced Databases)
- (Maggie, 3) x (2, Quantitative Risk Assessment)
- (Maggie, 3) x (4, Diversity Management)
- (Maggie, 3) x (3, Microeconomics)
- (Maggie, 3) x (1, Marketing and Public Relations)
- (John, 4) x (1, Advanced Databases)
- (John, 4) x (2, Advanced Databases)
- (John, 4) x (2, Quantitative Risk Assessment)
- (John, 4) x (4, Diversity Management)
- (John, 4) x (3, Microeconomics)
- (John, 4) x (1, Marketing and Public Relations)

BNL:

- (Peter, 1) x (1, Advanced Databases)
- (Peter, 1) x (2, Advanced Databases)
- (Kate, 2) x (1, Advanced Databases)
- (Kate, 2) x (2, Advanced Databases)
- (Peter, 1) x (2, Quantitative Risk Assessment)
- (Peter, 1) x (4, Diversity Management)
- (Kate, 2) x (2, Quantitative Risk Assessment)
- (Kate, 2) x (4, Diversity Management)
- (Peter, 1) x (3, Microeconomics)
- (Peter, 1) x (1, Marketing and Public Relations)
- (Kate, 2) x (3, Microeconomics)
- (Kate, 2) x (1, Marketing and Public Relations)
- (Maggie, 3) x (1, Advanced Databases)
- (Maggie, 3) x (2, Advanced Databases)
- (John, 4) x (1, Advanced Databases)
- (John, 4) x (2, Advanced Databases)
- (Maggie, 3) x (2, Quantitative Risk Assessment)

- (Maggie, 3) x (4, Diversity Management)
- (John, 4) x (2, Quantitative Risk Assessment)
- (John, 4) x (4, Diversity Management)
- (Maggie, 3) x (3, Microeconomics)
- (Maggie, 3) x (1, Marketing and Public Relations)
- (John, 4) x (3, Microeconomics)
- (John, 4) x (1, Marketing and Public Relations)

(2, Quantitative Risk Assessment) tuple is brought into memory

- **4** times in NLJ
- **2** times in BNL

2) **Merge-Join(MJ)**:

- sort R1 on the join attribute (a)
- sort R2 on the jon attribute (b)
- r1 ← get first row of R1
- r2 ← get first row of R2
- while not at the end of either relations
    - if r1 joins with r2 (r1.a == r2.b)
        - output (r1, r2)
    - else if r1.a < r2.b
        - r1 ← get next row of R1
    - else
        - r2 ← get next row of R2
- end

3) This question is an open-ended question because we can't exactly say System R does good estimate or not without knowing exact proportion of the matching tubles but here is the my reasoning:

- P.age > 90
    - System R uses (high key - value) / (high key - low key)

Therefore, a patient can be in any valid range, from 0 babies to 120 very old people, even if there is only one patient whose age is 120, high key is increased to 120 so selectivity factor for this boolean factor becomes (120 - 90) / (120 - 0) = ¼ in our assumption.

- P.numof_visits_last_year > 2
    - System R uses the same formula to calculate the selectivity factor.

Moreover, we can assume people usually visit their doctor once a month which gives us 12 as high key and we can assume that a patient have done at least one visit to the doctor to be his patient so selectivity factor will be (12 - 2) / (12 - 1) = 10 / 11

- P.age > 90 AND P.numof_visits_last_year > 2
    - will be calculated by F(P.age>90) * F(P.numof_visits_last_page>2)
    - so value is (1/4) * (10/11) = 10 / 44 ~ ¼

Here, visit numbers doesn't change age factor much since people generally go to the doctor 2 or more, its estimate nearly selects all tuples. However, patients older than 90 should be minority, not as much as ¼ of all patients so it doesn't seem so much reasonable.