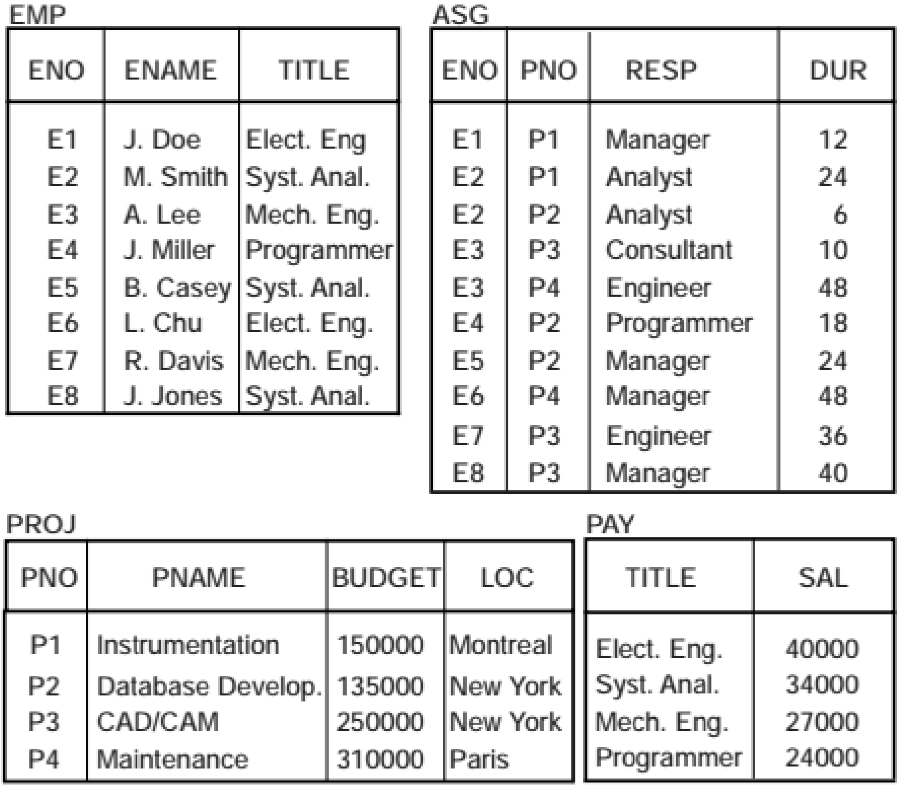
DISTRUBTED DB DESIGN

Exercise 01

Consider relation ASG in the Figure bellow. Suppose there are two applications that access ASG. The first is issued at five sites and attempts to find the duration of assignment of employees given their numbers. Assume that managers, consultants, engineers, and programmers are located at four different sites. The second application is issued at two sites where the employees with an assignment duration of less than 20 months are managed at one site, whereas those with longer duration are managed at a second site. Derive the primary horizontal fragmentation of ASG using the foregoing information.



EMP

ENO	ENAME	TITLE
E1	J. Doe	Elect. Eng
E2	M. Smith	Syst. Anal.
E3	A. Lee	Mech. Eng.
E4	J. Miller	Programmer
E5	B. Casey	Syst. Anal.
E6	L. Chu	Elect. Eng.
E7	R. Davis	Mech. Eng.
E8	J. Jones	Syst. Anal.

ASG

ENO	PNO	RESP	DUR
E1	P1	Manager	12
E2	P1	Analyst	24
E2	P2	Analyst	6
E3	P3	Consultant	10
E3	P4	Engineer	48
E4	P2	Programmer	18
E5	P2	Manager	24
E6	P4	Manager	48
E7	P3	Engineer	36
E8	P3	Manager	40

PROJ

PNO	PNAME	BUDGET	LOC
P1	Instrumentation	150000	Montreal
P2	Database Develop.	135000	New York
P3	CAD/CAM	250000	New York
P4	Maintenance	310000	Paris

PAY

TITLE	SAL
Elect. Eng.	40000
Syst. Anal.	34000
Mech. Eng.	27000
Programmer	24000

Exercise 02:

Consider table “EMP(EmpID, Name, Sal, Loc, Dept).”. There are four applications running against this table as shown below. Design a fragmentation strategy that satisfies the needs of these applications.

App1: “Select EmpID, Sal From EMP;”
App2: “Select EmpID, Name, Loc, Dept From EMP Where Dept = ‘Eng’;”
App3: “Select EmpID, Name, Loc, Dept From EMP Where Loc = ‘STP’;”
App4: “Select EmpID, Name, Loc, Dept from EMP Where Loc = ‘MPLS’;”

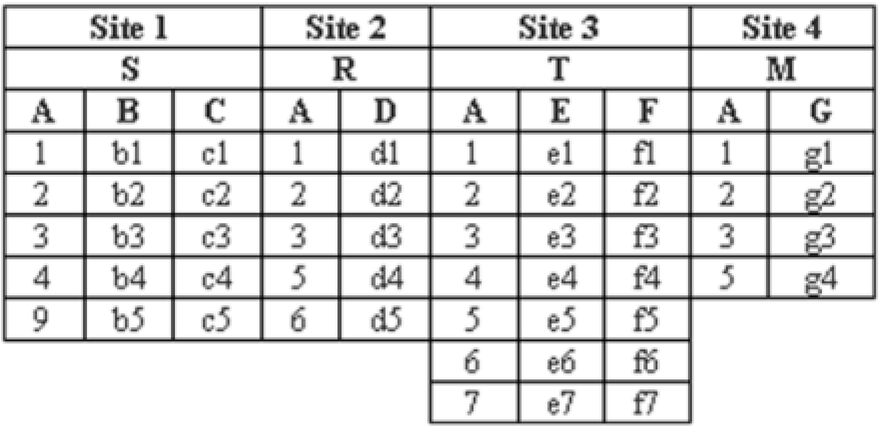
QUERY PROCESSING

Exercice 01

How many alternatives exist for joining three tables together? How many alternatives exist for joining four tables and five tables together? Can you extrapolate from these numbers to arrive at the number of alternatives that can be used to join *N* tables together?

Exercice 02

A four-site system has tables R, S, T, and M stored as depicted bellow. The user is at Site 1. Assume each row of any table that has more than one column can be sent in one message. You can also assume that a one-column table can be sent in one message. Also assume that the cost of sending each message is C units of time. There are no costs associated with sending commands. What is the cost of "(S JN R) JN (T JN M)" if "S JN R" and "T JN M" are done using semi-join and the final join is done normally? Make sure you show all steps and cost of communication for each step.

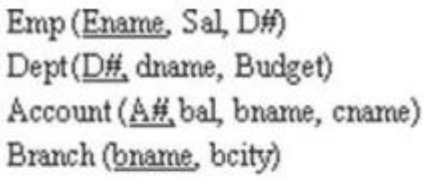


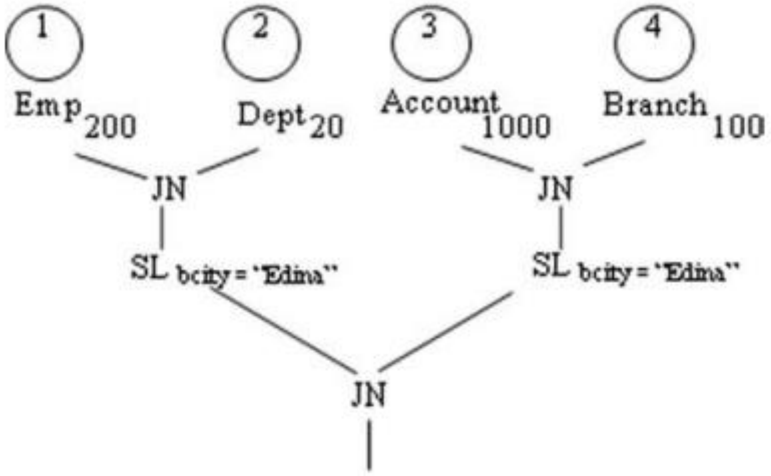
Site 1			Site 2		Site 3			Site 4	
S			R		T			M	
A	B	C	A	D	A	E	F	A	G
1	b1	c1	1	d1	1	e1	f1	1	g1
2	b2	c2	2	d2	2	e2	f2	2	g2
3	b3	c3	3	d3	3	e3	f3	3	g3
4	b4	c4	5	d4	4	e4	f4	5	g4
9	b5	c5	6	d5	5	e5	f5		
					6	e6	f6		
					7	e7	f7		

Exercice 03

Assume the four-site system shown in bellow—the table distribution and the query tree that prints all information about the employees who work for the engineering department and have an account in a branch in the city of Edina. Also assume the user is at Site 2. Let's assume that 40% of accounts are in branches in Edina; 50% of employees work for the engineering department; and 10% of employees who work for the engineering department have an account in a branch in Edina. If each row being sent from each site to any other site takes C units of time, what is the communication cost of the optimized query tree (results must be displayed to the user) in terms of C? Make sure you show individual step's cost and the total cost. Assume there is no cost associated with sending the commands.

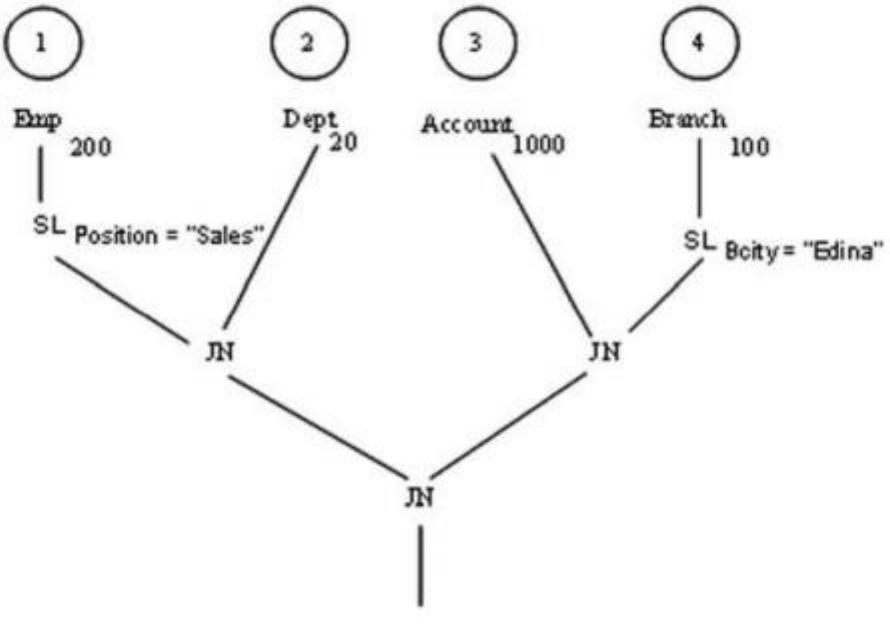
What is the cost of an optimal distributed execution strategy for this query in terms of the number of messages required? Assume the cost of sending each row of any table is C units of time. Assume there is no cost associated with sending the commands. Make sure you show all steps and the cost of each step.


Emp (Ename, Sal, D#)
Dept (D#, dname, Budget)
Account (A#, bal, bname, cname)
Branch (bname, bcity)



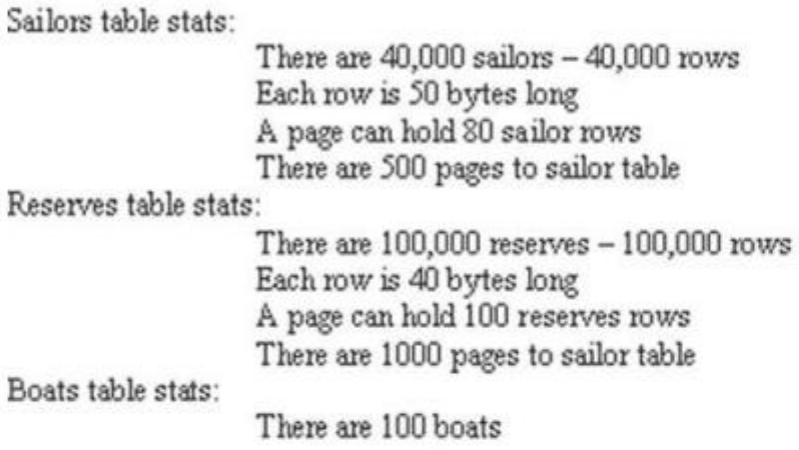
Exercice 04

In the four-site system shown in Figure 4.37, the user is at Site 1. We need to print all information about all sales persons who have accounts in a branch in the city of Edina. We know that 5% of branches are in Edina; 10% of all accounts are in branches in Edina; and 20% of employees are in sales. If processing each row takes “t” units of time, what is the processing cost of this tree? Make sure you show all steps and the cost of each step.



Exercice 05

We consider a database that has three tables as shown in the Figure bellow. Suppose we have a query that returns a sorted order of the sailors with a rating of 10 who have reserved the boat with ID = 50. The query plan for this question is depicted in the Figure. Assume the page size is 4000 usable bytes after the page overhead. If the first join is done by a nested-loop, row-based join and the second join is done by a sort–merge join, what is the cost of this plan in number of disk I/Os? Assume that we have five buffers for sorting and all distributions are uniform.



Sailors table stats:

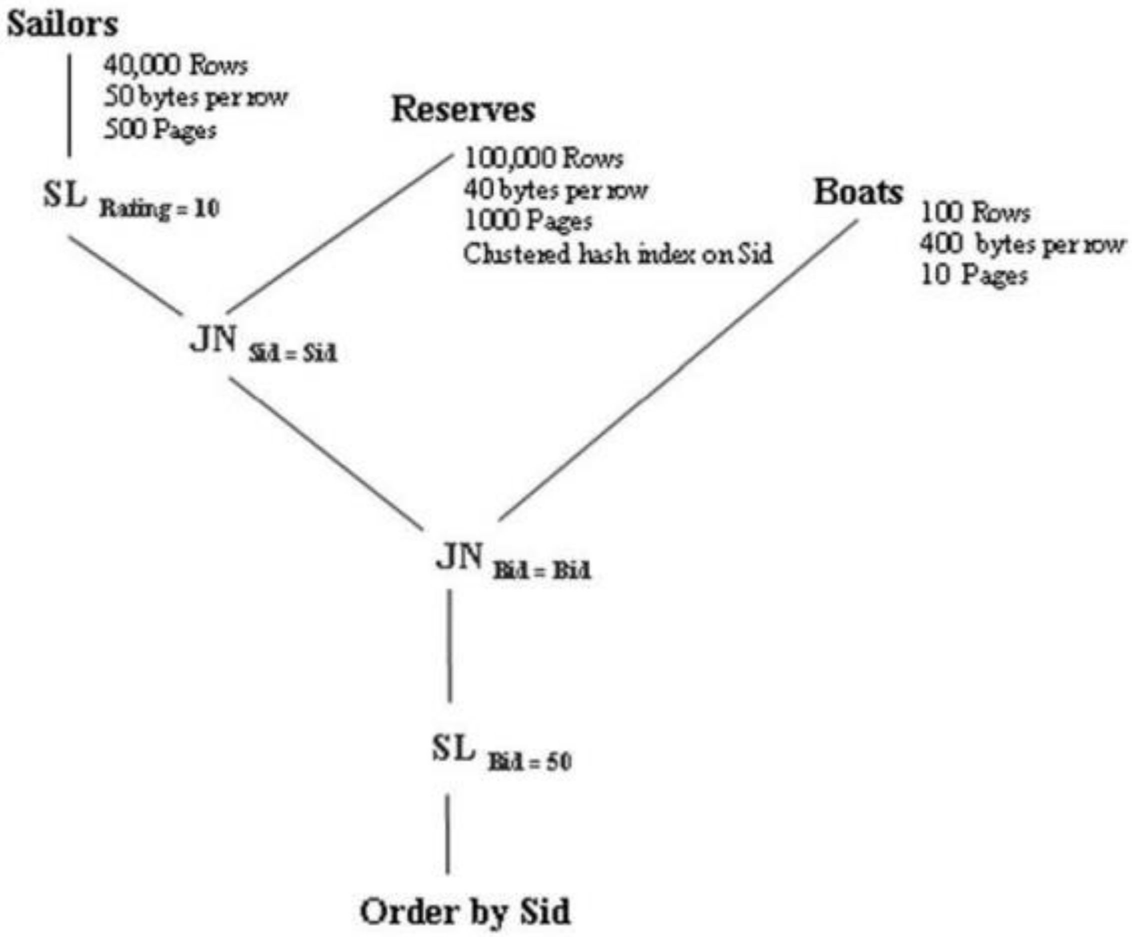
There are 40,000 sailors – 40,000 rows
Each row is 50 bytes long
A page can hold 80 sailor rows
There are 500 pages to sailor table

Reserves table stats:

There are 100,000 reserves – 100,000 rows
Each row is 40 bytes long
A page can hold 100 reserves rows
There are 1000 pages to sailor table

Boats table stats:

There are 100 boats



Exercise 06

A student library has a distributed database with relation ‘Students’ at Site S1 and a relation ‘OverDue’ at Site S2. Loans that are overdue are stored in the relation ‘Overdue’. S1:

Students : (StudentId, StudentStatus, M ajor). S2: OverDue: (OverDueId, StudentId, ItemId, DateDue, DateReturned).

The length of the attributes are 7 bytes for all attributes except StudentStatus which is 3 bytes.

Out of the total 24 000 students at the university, on the average 6 000 students will have at any given time, *ONE single* book (or some item) which is overdue. 10 percent of the students are majoring in Computer Science (‘CS’).

At site S3 we need to evaluate the following query to get those loans from the Computer Science majors which are overdue:

```
SELECT DISTINCT S.StudentId, S.Major, OD.DateReturned FROM Students S, OverDue OD
WHERE OD.StudentId = S.StudentId AND S.Major = ‘CS’ AND OD.DateReturned IS NULL
```

Estimate the cost of the query execution plan according to the number of bytes to be transferred?

Exercise 07

A student library has a distributed database with relation ‘Students’ at Site S1 and a relation ‘OverDue’ at Site S2. Loans that are overdue are stored in the relation ‘Overdue’. S1: Students : (StudentId, StudentStatus, M ajor). S2: OverDue: (OverDueId, StudentId, ItemId, DateDue, DateReturned).

The length of the attributes are 7 bytes for all attributes except StudentStatus which is 3 bytes.

Out of the total 24 000 students at the university, on the average 6 000 students will have at any given time, *ONE single* book (or some item) which is overdue. 10 percent of the students are majoring in Computer Science (‘CS’).

At site S3 we need to evaluate the following query to get those loans from the Computer Science majors which are overdue:

```
SELECT DISTINCT S.StudentId, S.Major, OD.DateReturned FROM Students S, OverDue OD
WHERE OD.StudentId = S.StudentId AND S.Major = ‘CS’ AND OD.DateReturned IS NULL
```

Estimate the cost of the optimized query execution plan according to the number of bytes to be transferred?

Exercise 08

A Banking DB contains the following relations among others:

Accounts (AccountNr, Balance),

DebitCards (CardNr, AccountNr, CustomerId),

Customers (CustomerId, FullName, Ssn, Address, City)

Transactions (Site, TransactDate, TransactType, Amount, AccountNr, CardNr),

Each transaction such as a debit or a query for the balance is recorded as a row in the relation Transactions. The relation is hash horizontally fragmented according to attribute AccountNr.

The relation DebitCards as well as the relation Transactions are hash horizontally fragmented using the attribute CardNr (Debitcard's number) using the same hashing function. The relation Customers is range- horizontally fragmented according to attribute FullName.

1. How do we compute the following SQL-clauses?

   a] select count(*)from Transactions where TransactDate > '2008-01-01'.

   b] select CustomerId from Customers where FullName = 'Smith K. John'.

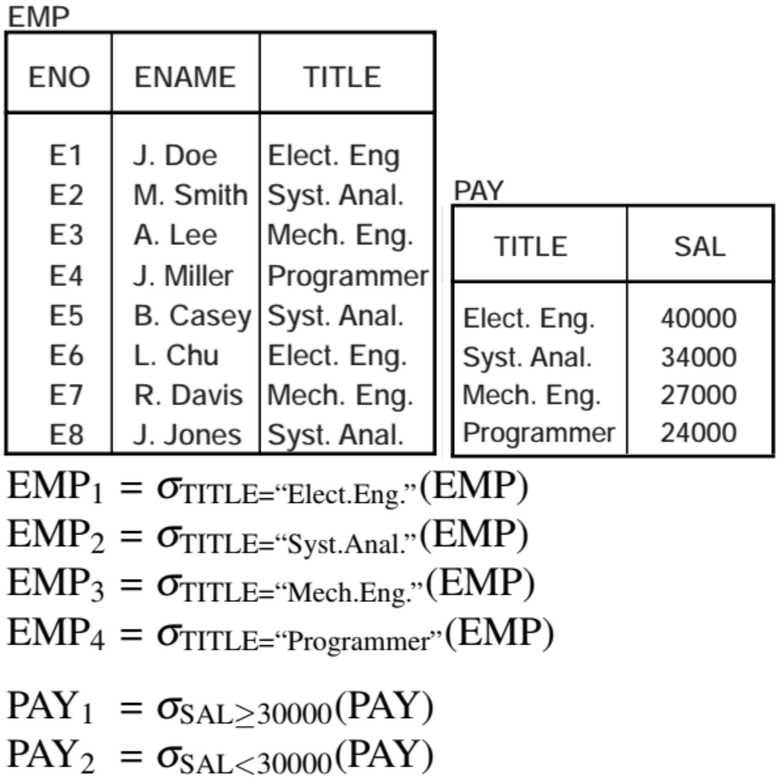
   c] update Accounts set balance = balance − 50 where AccountNr = 1234.

   d] insert into Transactions values('site1', '2008-01-01', 'WITHDRAW', 100, 1234, 5678).

2. Give the optimized plan execution of the query:

   select T.site, T.TransactDate, T.TransactType, T.amount, D.CustomerId from DebitCards D, Transactions T,Customers C
   where C.FullName = 'Smith, K. John' and C.CustomerId = D.CustomerId and D.CardNr = T.CardNr order by T.TransactDate

Exercice 09. Consider relations EMP and PAY in the Figure bellow. EMP and PAY are horizontally fragmented as indicated bellow. Draw the join graph of EMP JN PAY.



EMP

ENO	ENAME	TITLE
E1	J. Doe	Elect. Eng
E2	M. Smith	Syst. Anal.
E3	A. Lee	Mech. Eng.
E4	J. Miller	Programmer
E5	B. Casey	Syst. Anal.
E6	L. Chu	Elect. Eng.
E7	R. Davis	Mech. Eng.
E8	J. Jones	Syst. Anal.

PAY

TITLE	SAL
Elect. Eng.	40000
Syst. Anal.	34000
Mech. Eng.	27000
Programmer	24000

$$EMP_1 = \sigma_{TITLE=\text{“Elect.Eng.”}}(EMP)$$
$$EMP_2 = \sigma_{TITLE=\text{“Syst.Anal.”}}(EMP)$$
$$EMP_3 = \sigma_{TITLE=\text{“Mech.Eng.”}}(EMP)$$
$$EMP_4 = \sigma_{TITLE=\text{“Programmer”}}(EMP)$$

$$PAY_1 = \sigma_{SAL\geq 30000}(PAY)$$
$$PAY_2 = \sigma_{SAL<30000}(PAY)$$

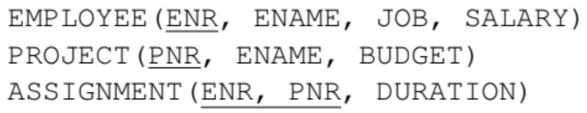
Exercice 10 Given the following relations and SQL query:
Student (sid, name, age, address)
Book(bid, title, author)
Checkout(sid, bid, date)

SELECT S.name FROM Student S, Book B, Checkout C
WHERE S.sid = C.sid AND B.bid = C.bid AND B.author = ’Olden Fames’ AND S.age > 12
AND S.age < 20

1) Write the optimized query execution plan
2) Assume that Student is in Site1, Book in Site2 and Checkout is in Site3, and the query is executed from Site1. Write the query execution plan.

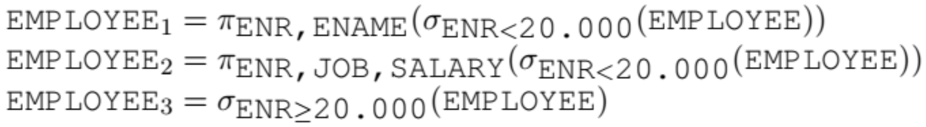
Exercice 10.

Given the following relational schema :



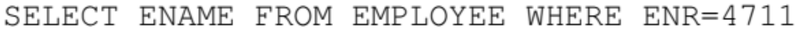
```
EMPLOYEE(ENR, ENAME, JOB, SALARY)
PROJECT(PNR, ENAME, BUDGET)
ASSIGNMENT(ENR, PNR, DURATION)
```

The relation EMPLOYEE is fragmented as follows:



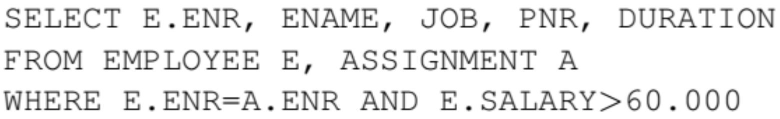
$$\mathrm{EMPLOYEE}_1 = \pi_{\mathrm{ENR,ENAME}}(\sigma_{\mathrm{ENR}<20.000}(\mathrm{EMPLOYEE}))$$
$$\mathrm{EMPLOYEE}_2 = \pi_{\mathrm{ENR,JOB,SALARY}}(\sigma_{\mathrm{ENR}<20.000}(\mathrm{EMPLOYEE}))$$
$$\mathrm{EMPLOYEE}_3 = \sigma_{\mathrm{ENR}\geq 20.000}(\mathrm{EMPLOYEE})$$

1) Perform algebraic optimization for the following query?



```
SELECT ENAME FROM EMPLOYEE WHERE ENR=4711
```

The following query on EMPLOYEE and ASSIGNMENT has to be processed:



```
SELECT E.ENR, ENAME, JOB, PNR, DURATION
FROM EMPLOYEE E, ASSIGNMENT A
WHERE E.ENR=A.ENR AND E.SALARY>60.000
```

Furthermore, the following statistics are available: card(EMPLOYEE) = 1.000, card(ASSIGNMENT) =.1.500; both relations are stored on different nodes. The query is initiated on a third node N and the result must be available there. The salary condition is satisfied by 20% of the employees; 25% of the employees do not work for any specific project.

2) What is the optimal join processing strategy ?