COLLEGE OF ENGINEERING PUNE

(An Autonomous Institute of Government of Maharashtra.)

END Semester Examination

(CT(DE)-14004) Advanced Database Management Systems

Course: B.Tech Branch: Information Technology /Computer Engineering

Semester: Sem VII

Year: 2014-2015 Max.Marks:60

Duration: 3 Hours Time:- 2-5 **Date:28/11/2014**

Instructions: **MIS No.** □□□□□□□□ □

1. Figures to the right indicate the full marks.
2. Mobile phones and programmable calculators are strictly prohibited.
3. Writing anything on question paper is not allowed.
4. Exchange/Sharing of anything like stationery, calculator is not allowed.
5. Assume suitable data if necessary.
6. Write your MIS Number on Question Paper

Q. 1 a) Specify a good way to parallelize 3

1. Difference operation
2. Aggregation by **count** operation
3. Aggregation by **avg** operation

b) Describe the benefits and drawbacks of Pipelined parallelism. 2

c) Consider a parallel DBMS in which each relation is stored by horizontally partitioning its tuples across all disks. 5

Employees (*eid*:integer, *did*:integer, *sal*:real)
Departments (*did*:integer, *mgrid*:integer, *budget*:real)

The *mgrid* field of Departments is the *eid* of the manager. Each relation contains 20-byte tuples, and the *sal* and *budget* fields both contain uniformly distributed values in the range 0 to 1,000,000. The Employees relation contains 100,000 pages, the Departments relation contains 5,000 pages, and each processor has 100 buffer pages of 4,000 bytes each.

The cost of one page I/O is t_d, and the cost of shipping one page is t_s; tuples are shipped in units of one page by waiting for a page to be filled before sending a message from processor i to processor j. There are no indexes, and all joins that are local to the processor are carried out using a sort-merge join. Assume that the relations are initially partitioned using a round-robin algorithm and that there are 100 processors.

For each of the following queries, describe the evaluation plan briefly and give its cost in terms of t_d and t_s. You should compute the total cost across all sites as well as the 'elapsed time' cost (i.e., if several operations are carried out concurrently, the time taken is the maximum over these operations).

1. Find the highest paid employee over all departments with budget < 300,000.
2. Find the salaries of all managers.
3. Find the highest paid employee over all departments with budget <100,000.

Q.2 a) Consider the Employees and Departments relations described in Question 1.c They are now stored in a distributed DBMS with all of Employees stored at Naples and all of Departments stored at Berlin. There are no indexes on these relations. The cost of various operations is as described in Question 1.c. Consider the query: 6

```
SELECT *  FROM Employees E, Departments D
                    WHERE E.eid = D.mgrid;
```

The query is posed at Delhi and you are told that only one percent of the employees are managers. Find the cost of answering this query using each of the following plans:

1. Compute the query at Naples by shipping Departments to Naples; the ship the result to Delhi.
2. Compute the query at Delhi by shipping both relations to Delhi.
3. Compute the query at Naples using Semijoin; then ship the result to Delhi.

OR

Histograms are used for constructing load –balanced range partitions. Suppose the histogram has the values between 1 and 100 and are partitioned into 10 ranges, 1-10, 11-20.........91-100, with frequencies 5,5,20,10,5,5,20,5 and 5 respectively. Give a load balanced range partitioning function to divide the values into 5 partitions

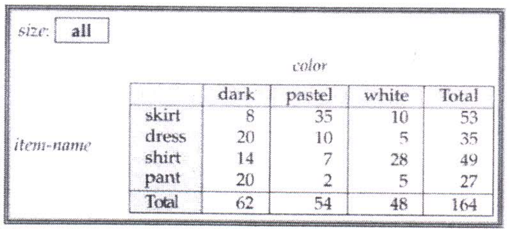
b) In context to distributed databases 4

1. Describe data transparency and its various forms.
2. Role of Multi database system and gateway in heterogeneous distributed databases.

Q. 3 a) Given that LDAP functionality can be implemented on the top of a database system, what is the need for the LDAP standard? 3

b) Describe OLAP in detail. 3

c) Consider the ***sales*** relation. Write an SQL query to compute the cube operation on the relation, giving the relation in figure. Do not use the **cube** construct. 4



size: all

color

item-name	dark	pastel	white	Total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pant	20	2	5	27
Total	62	54	48	164

Q.4 Make use of the following table, a hypothetical employment statistics record for Recent graduates.

ID	major	avg. project score	avg. exam score	co-op?	employed?	Salary
1	Computer	87	75	Y	Y	60,000
2	History	?	92	N	N	?
3	Computer	77	95	N	Y	50,000
4	Engineering	97	65	N	N	0
5	Engineering	84	75	Y	Y	40,000

a) Diagram a star schema containing employment statistics data for recent graduates. Based your ideas on the table given, but think of other information that might be relevant. In particular, assume that the warehouse contains data for multiple universities as well as multiple majors. 5

b) Describe how a data cube can be used to determine 5

1. Which university graduates earn highest salaries?
2. Which university is the best place to study computers?

Q.5 a) Give the DTD for an XML representation of the following nested-relational schema 4

Emp=(ename,ChildrenSet **setof**(Children),SkillsSet **setof**(Skills))
Children=(name,Birthday)
Birtday=(day,month,year)
Skills=(type,ExamsSet **setof**(Exams))
Exams=(year,city)

b) Write a query in XPath on the DTD above to list all skill types in *Emp*. 2

c) Write a note on XSLT 4

Q.6 a) What is Hadoop framework? 5

b) Explain the word count implementation via the Hadoop framework. 5