Tutorial 4

Exercise 1 – State ONE reason to use a distributed DBMS instead of distributed processing. Also indicate ONE example of functionality that a distributed DBMS should be able to provide.

Exercise 2 a) – Consider a simplified distributed database Company consisting of the two relations:

Employee(eId, fName, lName, dNo); 5500 tuples stored in London
Department(dNo, dName, mgr_eId); 100 tuples stored in Beijing

The cost of transferring data over the network is usually high. User **P** at the site in Paris needs to retrieve **ALL** employee names with their department names where the employee works; determine the relational algebra expression for this, which will result in 5500 tuples (each one being 8 bytes long).

Exercise 2 b) – Consider again the scenario of **Exercise 5** and now also assume the below:

Each tuple in the **Employee** and **Department** relations is 15 and 30 bytes, respectively. Then determine which of the TWO execution strategies is more economical to retrieve and transfer results:

1. Move both relations to Paris and process the query at the Paris site.
2. Move the **Employee** relation to the Beijing site, execute the join operation at the Beijing site, and send the query results to the Paris site.

Exercise 2 c) – Consider again the scenario of **Exercise 5** and now also assume the below:

Fragmentation is being applied to the **Company** database, such that:

1. there are only 2 departments with ids “001” and “002”, and there will be two separate applications managing the employees of each department;
2. there are 2 applications to manage the departments and their managers: departments’ list (with **dNo**, **dName**) and managers’ list (with **dNo**, **mgr_eId**).

For each case, identify the type of fragmentation that should be applied and write the relational algebra expression to achieve it.

Exercise 3 – Fill in the gaps labelled **A – I** in the text below, which refers to XML and related technologies. A list of words is given to help you, but some are not necessary.

XML is a meta-language that enables designers to create their own ____**A**_____ tags to provide functionality not available with ____**B**_____.

XML retains the key ____**C**_____ advantages of ____**D**_____, structure, and ____**E**_____.

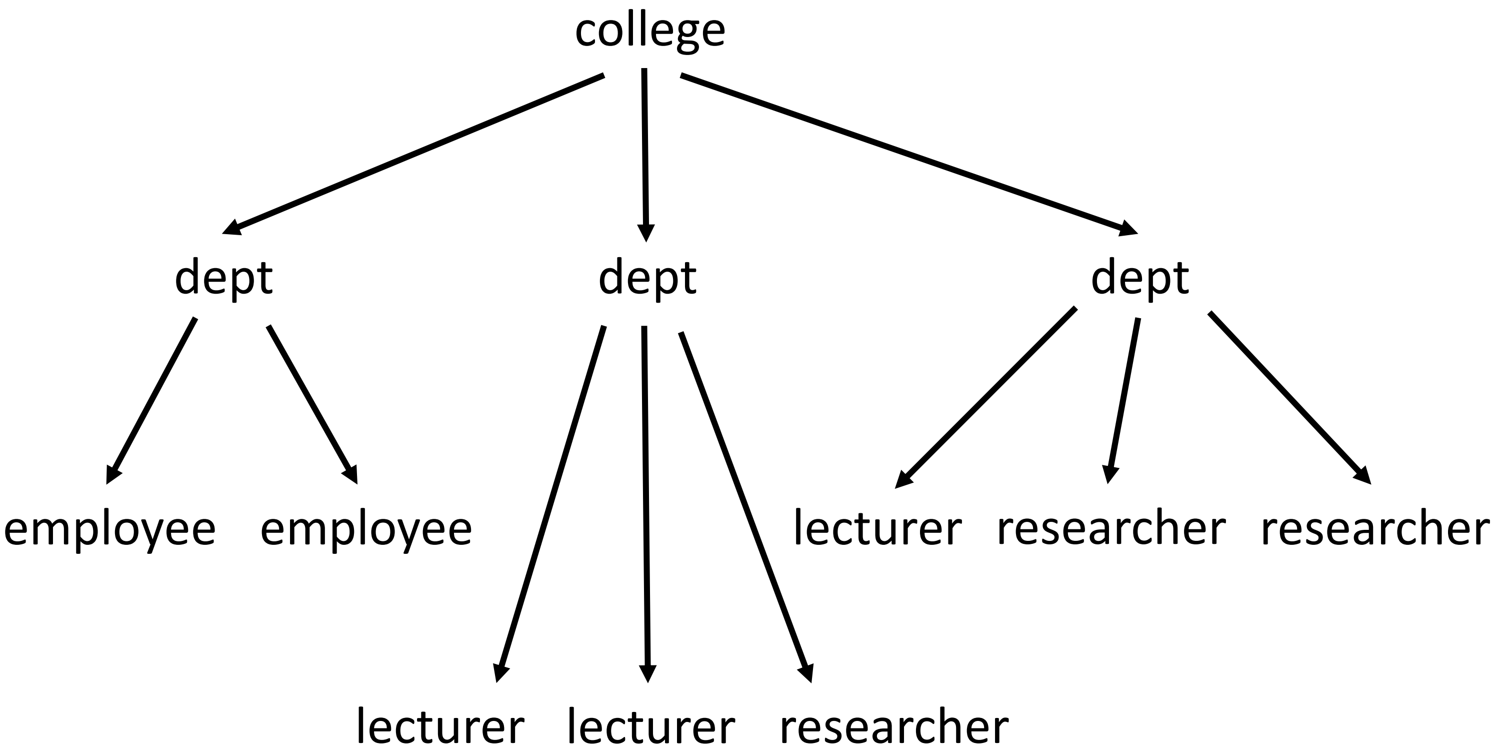
XML is designed to ____**F**_____ ____**G**_____ by enabling different kinds of data to be ____**H**_____ over the ____**I**_____.

- HTML
- HTML
- XSD
- DTD
- customised
- links
- extensibility

- SGML
- web
- exchanged
- XPath
- validation
- complement
- well-formed

Exercise 4 – Use the XML document and its corresponding tree representation, to then represent the data in relational format. You should include all the tables with names, columns and tuples.

```
<college>
   <dept name="Admin">
      <employee>John</employee>
      <employee>Dina</employee>
   </dept>
   <dept name="Engineering">
      <lecturer>Alma</lecturer>
      <lecturer>James</lecturer>
      <researcher>David</researcher>
   </dept>
   <dept name="Computing">
      <lecturer>Alina</lecturer>
      <researcher>James</researcher>
      <researcher>Alma</researcher>
   </dept>
</college>
```



Exercise 5 – Explain the concept of Data mining.

Exercise 6 – Identify all the CORRECT statements about the CAP theorem.

A. The A in the CAP theorem stands for Availability, which is about data sometimes not being available (e.g., if a server is down).

B. The CAP theorem assumes there are many nodes in the system, but the nodes don’t have replicas of partitions of the data.

C. Appropriate management of distributed data requires the 3 properties Consistency, Availability, and Partition tolerance.

D. When choosing a data model to store an organisation’s data, you should consider which of these properties are most important: CA, AP or CP.

Exercise 7 – Select the statements that CORRECTLY list differences between NoSQL and RDBMS systems.

A. NoSQL systems have looser schema definitions compared to RDBMSs.

B. NoSQL systems are not appropriate to handle distributed, large databases.

C. NoSQL systems come with a relaxation of the ACID properties.

D. NoSQL systems should be applied when frequent updates, as well as reads, are required.

E. Applications with very structured data and/or requiring high integrity and atomicity are better managed with a NoSQL system.

Exercise 8 – Fill in the gaps labelled **A – P** in the text below, which refers to NoSQL and RDBMS systems. A list of words is given to help you, but some are not necessary.

___**A**___ databases are non-relational data management systems that do not require a ___**B**___ schema and are ___**C**___ to scale; they are mainly aimed at distributed data stores with ___**D**___ ___**E**___ data storage needs. Therefore, ___**F**___ is used for ___**G**___ ___**H**___ and real-time web apps.

Traditional ___**I**___ use ___**J**___ ___**K**___ to store and retrieve ___**L**___ data. On the other hand, a ___**M**___ database system includes a range of database technologies that can store ___**N**___, ___**O**___ and ___**P**___ data.

large
NoSQL
small
NoSQL
RDBMSs
NoSQL
RDBMSs
fixed
loose
very
easy

structured
RDBMSs
SQL
big
semi-structured
data
syntax
structured
semi-structured
difficult
unstructured