Assignment 2
Data Management for the Big Data Age

1 Assignment Description

This is a group assignment. Students are referred to the University's policy on plagiarism. The aim of this assignment is to develop skills to use XML (and JSON) to design and query semi-structured data, and to evaluate the difference between document-based NoSQL databases and SQL databases.

2 Task 1: XML and DTD

A relational database for a company is described below:

> The database is part of a company information system and contains four tables; *Division* describing the divisions of the company, *Project* describing the projects within the divisions, *Employee* containing information about employees within the divisions, and *Assign* describing employees' involvement in projects. The schema for each of the tables is shown below, with the primary key attributes underlined, the *foreign key attributes* in italic font, and some more information about each table.

Division	(DID, DNAME, LOCATION)
	The company has several divisions identified by DID. Each division has a name (DNAME) and a location (LOCATION)
Project	(PID, PNAME, BUDGET, *DID*)
	Each project has an identifier PID, a name (PNAME), a budget (BUDGET) and belongs to a division which is in the Division table.
Employee	(EID, ENAME, OFFICE, BIRTHDATE, SALARY, *DID*)
	Each employee has an identifier EID, a name (ENAME), an office room (OFFICE), a birth date (BIRTHDATE), a salary (SALARY) and belongs to a division which is in the Division table.
Assign	(*PID, EID*, HOURS)
	Each employee may be assigned to one or more projects that belong to the same division as the employee. Each project has one or more employees.

The number of hours is recorded.

The domains for each of the attributes are:

Division

DID	3 character string	Unique identifier for the division.
DNAME	Up to 20 characters	Unique name of the division.
LOCATION	Up to 30 characters	Location of the division.

Project

PID	4 character string	Unique identifier for the project.
PNAME	Up to 40 characters	Name of the project.
BUDGET	8 digit number	Budget of the project.
DID	3 character string	Identifies the division which the project belongs to.

Employee

EID	6 character string	Unique identifier for the employee.
ENAME	Up to 30 characters	Name for the employee.
OFFICE	Up to 6 characters	Office room number.
BIRTHDATE	Date	Birth date of the employee.
SALARY	6 digit number	Salary of the employee.
DID	3 character string	Identifies the division which the employee belongs to.

Assign

PID	4 character string	Identifies the project.
EID	6 character string	Identifies the employee.
HOURS	4 digit number	Hours the employee has worked for the project.

You are required

2.1 Design a proper DTD called company.dtd for the information given by the above relational database schema. The valid XML documents under this DTD must have a tree structure with **as much nesting as possible**. The DTD must also capture all the primary key and foreign key constraints.

2.2 Populate an XML document called company.xml by referring company.dtd. It has at least 2 divisions. Each division has at least 2 projects and 3 employees. Each project has at least 2 employees in the same division working on it, and each employee works for at least 1 project. You are required to use company.dtd to validate company.xml. The domains for all attributes described above can be used to input values of attributes.

3 Task 2: XPath and JSON

3.1 For the XML document classes.xml shown below, write the XPath expressions for the following queries.

a) Find the instructor of all classes.
b) Find the title for those classes with “Grant” as an instructor.
c) Find the classID attribute for those classes with “yes” in attribute *req* of credits.
d) Find titles for those classes with more than one instructor.

3.2 Convert the XML document classes.xml to its JSON format.

classes.xml

```
<?xml version="1.0"?>
<classes>
   <class classID="CS115">
     <department>ComputerScience</department>
     <instructors>
       <instructor>Adams</instructor>
       <instructor>Dykes</instructor>
     </instructors>
     <title>Programming Concepts</title>
     <credits req="yes">3</credits>
   </class>
   <class classID="CS205" semester="fall">
     <department>ComputerScience</department>
     <instructors>
       <instructor>Grant</instructor>
     </instructors>
     <title>JavaScript</title>
     <credits req="yes">3</credits>
  </class>
  <class classID="CS255" semester="fall">
    <department>ComputerScience</department>
    <instructors>
      <instructor>Adams</instructor>
      <instructor>Grant</instructor>
       <instructor>Dykes</instructor>
    </instructors>
    <title>Java</title>
    <credits req="no">3</credits>
  </class>
</classes>
```

4 Submission requirements

You are asked to submit a single ZIP pdf file with the name Assignment2.zip to canvas with the following three files.

File name	Description
company.dtd	DTD for the company database (2.1)
company.xml	Well-formed XML document, must refer to company.dtd (2.2)
classes.pdf	XPath expressions for 3.1.a, 3.1.b, 3.1.c and 3.1.d and JSON representation for 3.2

For each submitted file, you are required to include the information about your group, including each member of the group. For each group, only one submission is needed.

You should use the provided Apache *Xerces validate* tool to validate the XML files. The command line is “validate company.xml”.

5 Marking Scheme

Work will be assessed based on the quality and presentation. The assignment will be marked out of 60 and will contribute **15%** towards assessment of the unit.

Assessment item	Marks
2.1 right nested structure (12) right info for Division/Project/Employee/Assign (4x3 = 12)	24
2.2 refer/follow DTD right (6) populate the XML document following the requirement (6)	12
3.1 a)/b)/c)/d) (4x4 = 16)	16
3.2 right JSON structure (4) right mapping (4)	8
Total	60