Reg. No.										

MANIPAL ACADEMY OF HIGHER EDUCATION

FIRST TERM POST GRADUATE DIPLOMA IN DATA SCIENCE (FULL TIME) DEGREE EXAMINATION – OCTOBER 2018

SUBJECT: PDS 405 – DATA SCRAPING AND WRANGLING

Monday, October 15, 2018

Time: 09:30 – 12:30 Hrs. Max. Marks: 100

1A. Which of the following selector is used to identify all li tags with a attribute type with which contains a value "tuple":

i) li[type="tuple"]
ii) li.type="tuple"
iii) li.tuple
iv) li.type

1B. Which of the following tag is used to create hyperlinks in HTML?

i) li
ii) a
iii) h1
iv) img

1C. Which of the following command is used to read a html page through requests module?

i) page = requests.get('http://www.google.com').json
ii) page = requests.get('http://www.google.com').text
iii) page = requests.get('http://www.google.com').html
iv) page = requests.get('http://www.google.com')

1D. Three DML commands of SQL are:

i) CREATE, ALTER, DELETE
ii) INSERT, UPDATE, DELETE
iii) CREATE, UPDATE, DROP
iv) CREATE, ALTER, DROP

1E. Which statement is true regarding the default behavior of the ORDER BY clause in SQL?

i) In a character sort, the values are case-sensitive
ii) NULL values are not considered at all by the sort operation
iii) Only those columns that are specified in the SELECT list can be used in the ORDER BY clause
iv) Numeric values are displayed from the maximum to the minimum value if they have decimal positions

1F. What is the output of the below query?

```
Select name, deptid
From employees
Order by salary
```

i) Sorts the data based on salary and lists name and deptid
ii) Sorts the data based on salary and lists name, salary and deptid
iii) Sorts the data based on name and lists name and deptid
iv) Error

1G. Which of these statements is TRUE regarding constraints in a relation?

i) ✓ A primary key is a candidate key
ii) All candidate keys are primary keys
iii) A foreign key should be a unique in its parent relation
iv) A foreign key cannot have null values

1H. Which of the below best defines the Data Definition Language (DDL) commands of SQL?

i) Describes how data is structured in the data base
ii) Provides a mechanism to alter the structure of the data
iii) Provides a mechanism to remove the structure of the data permanently
iv) All of the above

1I. IN SQL If where clause restricts rows which clause of select restricts groups:

i) Partition by
ii) Having
iii) Order by
iv) Over

1J. What is the significance of the statement "HAVING COUNT (emp_id)>2" in the given query?

```
SELECT d.name, count (emp_id) as emp_count
FROM department d INNER JOIN Employee e
ON d.dept_id=e.emp_id
GROUP BY d.name
HAVING COUNT (emp_id)>2
```

i) Filter out all rows whose total emp_id below2
ii) Selecting those departments having total number of employees > 2
iii) Both (i) and (ii)
iv) None of these

(2 marks × 10 = 20 marks)

2A. With a good example explain what the advantage is of using JSON data structure compared to tabular structure.

2B. What is the difference between scraping data from an API vs scraping data from a web page?

2C. Write the regular expression patter for the following:

i) Retain only alphabets, numbers and spaces in string.
ii) Retain only numbers with 10 digits.

2D. Assume you have employees and departments tables with the below structure:

Emp(empid, empname, salary, deptid)
Dept(deptid, dname, location_id)

Write a select query to:

i) Find out the department name and average salary of employees in each department.
ii) List Department name and no. of employees working in them.

2E. Write a SELECT query to list all types of CLERKs who draw the same salary. (Assume Employee table structure: Employee_id, Employee_Name, Dept_id, salary, Job_id, Manager_id)

2F. Consider the employees table:

(Assume Employee table structure: Employee_id, Employee_Name, Dept_id, salary, Job_id, Manager_id)

Identity and Fix the errors(if any) the below queries:

i) Select Employee_name, salary*12 as "Annual Salary"
From employees
Where "Annual Salary" > 12000

ii) Select last_name, job_id
From employees
Where salary > 1200
Order by salary desc

2G. What is Transaction Control Language? List the TCL commands in SQL and identify the purpose of each.

2H. List any two differences between single row subquery and a multiple row subquery. Give an example for each by detailing the operators used in each case.

2I. The below queries result in an error when executed. Explain why and suggest a solution to fix the error.

SELECT department_id, SUM(salary) FROM Employees WHERE SUM(salary) > 1000 GROUP BY department_id;

SELECT department_id, job_id, SUM(salary) FROM Employees GROUP BY department_id;

2J. Write a SELECT query to list employees who are subordinates of 'John'. (Assume Employee table structure: Employee_id, Employee_Name, Dept_id, Job_id, Manager_id)

(4 marks × 10 = 40 marks)

3. Assume the following sample HTML content, from which we have to scrape data.

```
<div class="products">
    <div class="product" type="mobiles">
        <a href="/moto.html" class="product-link">
            <div class="title">Moto G5</div>
            <img src="/moto.png" />
            <div class="desc">
                <p>8GB RAM, 5inch display, 12MP Front camera</p>
            </div>
            <div class="category" type="moto"></div>
        </a>
    </div>
    <div class="product" type="mobiles">
        <a href="/apple.html" class="product-link">
            <div class="title">Iphone 6</div>
            <img src="/iphone.png" />
            <div class="desc">
                <p>8GB RAM, 4inch display, 8MP Front camera</p>
            </div>
            <div class="category" type="apple"></div>
        </a>
    </div>
</div>
```

Write a code in R or Python to get the following data frame:

Title	URL	Image URL	Description	Category
Moto G5	/moto.html	/moto.png	8GB RAM,…	moto
Iphone 6	/apple.html	/iphone.png	8GB RAM,…	apply

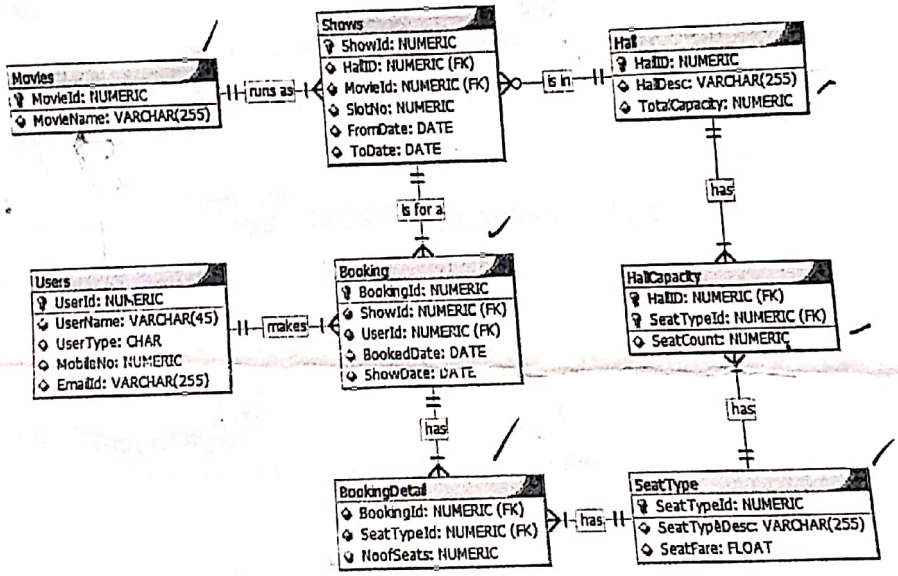
4. Explain any 5 aggregate functions in SQL with examples.

5. What are the different types of SQL statements? Explain each one of them briefly with the purpose of each.

6. The DB of the multiplex booking system has the following tables:

BOOKING (BOOKINGID,SHOWID,USERID,BOOKEDDATE,SHOWDATE)
BOOKINGDETAIL (BOOKINGID,SEATTYPEID,NOOFSEATS)
HALL (HALLID,HALLDESC,TOTALCAPACITY)
HALLCAPACITY (HALLID,SEATTYPEID,SEATCOUNT)
SEATTYPE (SEATTYPEID,SEATTYPEDESC,SEATFARE)
SHOWS (SHOWID,HALLID,MOVIEID,SLOTNO,FROMDATE,TODATE)
USERS (USERID,USERNAME,USERTYPE,MOBILENO,EMAILID)
MOVIES (MOVIEID,MOVIENAME)

Below is the relationship between the tables:



Write the below queries to find:

i) Show utilization of all halls for show date 15-Feb-2015 (no. of occupied seats vs. total seats)

ii) Show seat type-wise total booking (i.e. total no. of seats) and earnings for movie SHOLAY.

iii) List the total earnings from all movies for year 2015.

(10 marks × 4 = 40 marks)