**STW220CT: Data and Information Retrieval**

**Coursework – 2**

**Total Marks: 100**

**This is your STW220CT coursework 2. There are 3 tasks in this Coursework and all tasks are mandatory. All submissions through Moodle.**

**Task 1: Development of Graph database for a given dataset (40 Marks)**

Download “**ipl\_matches\_2008\_to\_2019.csv**” and “**match\_details.csv**” from Moodle which contains the details of IPL (Indian Premier League) from 2008 to 2019. You are expected to design and create a Graph database to visualize the dataset and to answer the following queries

1. Create a Data model diagram for the given datasets [5]
2. Create nodes and relationship according to the Data Model which you have created in question no 1. [5]
3. Show all the teams from season 2019. [1]
4. Show how many matches “Chennai Super kings” won and lost [3]
5. Show how many times does “MS Dhoni” got the player of the match when their teams won the match. [3]
6. How many matches does “Kings XI Punjab” won against “Mumbai Indians” when they lost the toss? [3]
7. Show all the matches in which “Chennai Super kings” won the toss, elected to bat and won the match. [3]
8. Show all the umpire and the number of matches they have umpired. [3]
9. Who umpired the most matches? [3]
10. Which team won most matches in “Wankhede Stadium” venue when they won the toss and chose to field in year 2008. [3]
11. Display top 10 batsman in IPL season 2008-2019. [3]
12. Delete all the matches from season 2008 except the matches played by “Mumbai Indians”.

[3]

1. Delete all the details of CH Gayle. [2]

**Task 2: Your Big Data Idea (30 marks)**

Identify and implement an idea that you have about how you would use Big Data for something intriguing.

1. Purpose an idea and clearly outlined (What is the purpose of your data collection and analysis). In the lecture notes above, you can see that each idea is specific and has a specific purpose.

2. Acquire the Data. You can do that in many ways including using available public large data sets.

3. Analyze the data in order to achieve the objective you set out for yourself in step 1.

4. Produce a report the includes your results, data visualization and thoughts.

**Evidence:**

Write a short report about what you did and how it worked out. No more than 1000 words. You should carry out research into these areas and reference your work using the CU Harvard Style.

**Task 3: Business to Business Portal (30 Marks)**

**Aim:**

In this task, you will design a back-end for a B2B (Business to Business eCommerce) portal’s website.

**Specs:**

The portal offers a variety of products for business users, each of which has specific qualities, some better than others. However, at this stage there is no total number of products, full list or characteristics (**hint:that sentence was important**)

To use the portal, customers must register a corporate account with the details stored. Moreover, the regular things you would expect from a portal like this should be available to the customer, this includes but is not limited to purchase history, wish lists and shopping carts.

Do not forget, this is a corporate account which means there will be more than one payment method used by customer, more than one billing address and more than one delivery address. Each one of the customers does have a username and password (In reality these would be encrypted but let’s assume they are not for this task).

According to the customers historical data, which for this tutorial are limited to items purchased and items searched for, recommendations are generated for the customer after logging in.

Your task:

**Part 1:**

Produce a complete design of the back-end data storage solution required. Things you have to consider

1. What type of queries will need to be done?
2. How fast do you need the results back?
3. How will you link the customer to the products bought?
4. Which fields should you use for comparison for recommendations?
5. What objects would be required? What documents?
6. Are the documents atomic or are there embedded documents?
7. Is this all a separate collection or do you want your whole database available for real time querying?

**Part 2:**

1. Produce diagrams explaining your solution.
2. Produce a list of queries you think might happen.