# ID2221 Data Intensive Computing (27 Oct., 2022)

1. **(2 points)** Assume you are designing a NoSQL database to store students' profiles. Each student has a unique ID and some extra information, which are not fixed among students. What data model do you use in your database? Moreover, assume you want to use consistent hashing to store students' profiles across computers. Explain how you can use consistent hashing to spread data in your network.

---

2. **(2 points)** Imagine you are working in a big company, and your company is planning to launch the next big Blogging platform. Tomorrow morning you go to your office and see the following mail from your CEO regarding a new work. How do you answer this email? Hint: use MapReduce to solve it, and explain what you do in Map and Reduce phases.

   ```
   As you know, we are building a blogging platform, and I need some statistics.  I need to find out, across all
   blogs ever written on our blogging system, how many times one-character words occur (e.g., "a", "I"), How many
   times two-character words occur (e.g., "be", "is"), and so on.  I know it's a really big job.  I am going on
   vacation for one week, and I must have this when I return.  Good luck.
   ```

---

3. **(2 points)** Explain the difference between the Transformations and Actions in Spark. Moreover, explain how Spark handles failures.

---

4. **(1 points)** Please list the differences between stateful and stateless operations in streaming processing programming models. Give two examples of each operation.

---

5. **(1 points)** What type of process (ETL or ELT) is executed in Dataware house and Data lake? Briefly explain their differences.

---

6. **(2 points)** In an undirected graph, each of two directly connected nodes is called *neighbors*. Assume you have an undirected graph in which each node stores a pair `(id, #neighbors)`, where the first item is the node's ID, and the second one is the number of node's neighbors. Write three pseudo-codes in the Pregel, Graphlab, and PowerGraph to find the node ID with the smallest number of neighbors. If two nodes have the same number of neighbors, choose the one with the smallest node ID.