



CentraleSupélec

CENTRALESUPÉLEC  
2022-2023

---

Big Data, Techniques and Platforms

## Document Databases

---

For this assignment you are required to use MongoDB. The only necessary file is:

- `recipes.json`

This file contains a set of recipes.

Download the files and store them in a directory on your computer. I suggest you to create, also in this case, a directory called `data`.

### 1 IMPORT DATA

Import data in your MongoDB database, and take advantage of the Studio3T graphical interface.

### 2 QUERYING DATA

Provide the queries that give you the following informations and the returned result.

1. **(1 point)** - Retrieve one element of the collection and provide a short description of the element. List if you see nested documents, lists, or any other attribute you consider worth to be commented.
2. **(1 point)** - Retrieve the name of the recipes (`Name` field) and the total time of preparation (`Total Time`).
3. **(1 point)** - Show the list of recipes (just their name) that contain the word "Lemon" in the name.

4. **(1 point)** - Show the number of recipes that are returned by the previous query.
5. **(1 point)** - Provide the name of the recipes that have Preparation Time greater than 40 minutes. Comment if you see something in the data representation that makes this operation not easy to be run.
6. **(1 point)** - Show the recipes with the lowest number of “servings”.
7. **(1 point)** - Look at data and provide an explication for the existence of both the attributes “Yield” and “Servings”.
8. **(1 point)** - Give the recipes (if any) that do not have ingredients listed.
9. **(3 point)** - Give the list of recipes that have at list 5 distinct ingredients listed. Comment the criteria that you used for this distinct operation.

**Important:** For this query and for any following query that needs to use the name of the ingredients you do not have to manipulate the names of the ingredients. If you have an ingredient whose name is “salt and pepper to taste” you do not have to parse and distinguish “salt” and “pepper” but just keep the name “salt and pepper to taste”.

10. **(3 points)** - Show the complete set of ingredients that appear in the database: show just the list of names.
11. **(1 point)** - Comment on how the ingredients are represented in the database and on how it is possible to query this field.
12. **(2 points)** - Find and show for each recipe 1) the number of steps (each step is an element of the Direction field) its preparation requires (the output must show the name of the recipe and the number of steps) 2) the name of the recipe.
13. **(3 points)** Give the number of recipes by ingredient.

### 3 BONUS - ENJOY WITH DATA

Now you can give freedom to your self-initiative and fantasy.

Using Spark, Dataframes and/or MongoDB provide one analysis of your choice on the dataset.