DS 5110: Introduction to Data Management and Processing (Fall 2022) Dr. Roi Yehoshua
Student name: (Due) November 25, 2022

MongoDB and Big Data

1 MongoDB

Download the file restaurants.json from Canvas and load it into a MongoDB database. The file contains 3772 documents.

Write the following MongoDB queries on the restaurants collection (you can write them either directly in the MongoDB shell or inside a python script):

1. Display all the restaurants located in the boroughs Bronx or Brooklyn.
2. Find the restaurant id, name, borough and cuisine for those restaurants whose name starts with the letters 'Mad'.
3. Find the restaurants that have received a score between 80 and 90.
4. Display the restaurant id and name of restaurants which have received 'C' grade in year 2014.
5. Find how many restaurants belong to each cuisine (note that the cuisine attribute may contain more than one cuisine type, e.g., "Ice Cream, Gelato, Yogurt, Ices").
6. Find the restaurants that do not prepare any cuisine of 'American' and their average grade score is higher than 30. Display the restaurant ids and their average score.
7. For each restaurant display only the grades that were recorded from the year 2014 onwards.
8. Calculate the average score across all the restaurants in the collection.

2 MapReduce

Describe how to implement the following relational operations using MapReduce. Write the map and reduce functions in pseudocode.

1. Projection $\pi_S(R)$: From each tuple of relation R produce only the components for the attributes in S.

2. Intersection $R \cap S$: Return the tuples that are present in both relations R and S. Assume that relations R and S have the same schema (same attributes and same type).

3. Grouping $\gamma_{A,\theta(B)}(R)$. Given a relation $R(A, B, C)$, with one grouping attribute A, one aggregated attribute B, and another attribute C, which is neither grouped or aggregated:

   (a) Partition the tuples of R according to their values in attribute A.

   (b) For each group, aggregate the values in attribute B and apply function θ on the aggregated value (θ is an aggregation operation such as SUM, COUNT or MAX).

   The result of this operation is one tuple for each group. That tuple has a component for the grouping attribute A, with the value common to tuples of that group. It also has a component for each aggregation $\theta(B)$, with the aggregated value for that group.

3 Spark

The bombing campaigns of the Vietnam War were the longest and heaviest aerial bombardment in history. The following datasets describe all the air force operations during the Vietnam War.

The file `Bombing_Operations.json` contains the following attributes:

- AirCraft: Aircraft model (example: EC-47)
- ContryFlyingMission: Country
- MissionDate: Date of the mission
- OperationSupported: Title of the operation (example: Operation Rolling Thunder)
- PeriodOfDay: Day or night
- TakeoffLocation: Take off airport
- TimeOnTarget
- WeaponType
- WeaponsLoadedWeight

The file `Aircraft_Glossary.json` contains the following attributes:

- AirCraft: Aircraft model (example: EC-47)
- AirCraftName
- AirCraftType

Load these datasets into Spark and answer the following questions:

1. Show the total number of missions for each of the countries involved (according to ContryFlyingMission). Write this query (a) With the DataFrame API (b) using Spark SQL (c) with RDD operations. Which of these methods was the most efficient?

2. Plot a bar chart with the number of missions by country.

3. Plot the number of missions per day for each of the countries involved.

4. How many takeoffs were launched to attack North Vietnam on 29 June 1966 from each location?

5. Which month saw the highest number of missions?

6. Which campaigns saw the heaviest bombings?

7. What was the most used aircraft type during the war (in terms of number of missions)?