1. This project encompasses a basic data pipeline, i.e., data retrieval and convenient accessibility to it.
2. In order to implement the above, MongoDB is chosen to be the database and MongoDB Atlas as its cloud equivalent.
3. At first, a client is established at the beginning of each part of the assignment in order to allow data retrieval from the cloud.
4. The "db\_setup" script generates a list of 20k meaningful strings, and then injects it to a new document named "records" (stored in a collection named "advertisements".
5. The script "sort\_program1" demonstrates a simple data retrieval, sorts it remotely and injects it to a new collection (named RESULTS) to the relevant cluster.
6. The script "sort\_program2" demonstrates limited data retrieval by limiting each query to 2000 strings.
   1. At first, the "data\_retrieval" function is used in order to manage the queries and store them into a designated list.
   2. Then, a "k\_way\_merge\_sort" is used in order to independently sort each list, and merge them to a single sorted list (see documentation inside the "Functions" script, which servers as a functions library).
7. **NOTE:** the access to the data base is given during the client setup phase (part 3 of this document). Everything inside it is accessible, assuming the names and nesting order of the required collections and documents is known. The Nesting is described as follows:

|  |  |  |
| --- | --- | --- |
| Database | Collection | Document |
| Advertisements | RESULTS | Sorting-step1 |
| Sorting\_step1\_Process\_time |
| Sorting\_step2\_Process\_time |
| records | 20k strings |