# FIT3182 2023 Assignment - Frequently Asked Questions

Last updated: 02 May 2024

**Q1:**

**Both Parts A and B**

Does it matter whether both parts of the assignment use:

a) Same collection as Part A, that is Part B will use the same collection for inserting, or

b) Different collections as Part A, that is Part B will use a new collection for inserting, or

c) Both are acceptable (same/different collection are acceptable)

Given that the specification did not mention explicitly on the relationship of the collections between Part A and B.

**Answer: You could use a different collection for Part B of the assignment to store results of the streamed climate and hotspot data (i.e., if the streamed hotspot coordinate matches the climate data coordinate). However, please make sure that you use the same data model for the collections in Part A and B of the assignment. In any case, if you are using a single collection for both Parts A and B of the assignment, that should be okay as well.**

**Q2:**

**Part A Task 2**

It is stated that “Please use a csv library to read the files.” Does this statement mean that:

a) In absolute terms, can use csv library ONLY

b) In abstract terms, can use csv library and all other libraries such as pandas

**Answer: In the assignment specification, we did ask students to use a Python csv library when reading from the historic csv files. However, I have received several requests from students to use pandas for Part A (and Part B) of the assignment, to which I have allowed them to do so. Therefore, you have the option to use either**

**the csv library or pandas. However, please take note of the pandas version to be used in a virtual machine or your native Linux setup..**

**Q3:**

**Part B Task 1(d)**

1. Does the climate data require storing of coordinates (latitude and longitude) given that each climate data has already mapped to the corresponding hotspots (which already contains the latitude and longitude)?

2. Will we merge hotspots when more than two hotspots all come from the same satellite?

3. In the case of the hotspots where they are to be merged, which latitude and longitude values are we going to store in the collection? There are some minor differences between these values after precision 5. Can we take just the first coordinate or any coordinate is acceptable?

4. If the climate record has no hotspots, should we still store the cause of fire (either natural or other) even when there are no fires?

**Answer:**

**1. True, it is not necessary to write the coordinates of the climate data into the collection because the hotspot data already contains the coordinates which is mapped to the climate data. Nevertheless, if you have already written it, I don't see any risks or issues here. If you do identify any issues, please inform your tutor.**

**2. This has been outlined in the assignment specs. To reiterate, you will merge all hotspots (regardless of their satellite origin) as long as they share a similar location and created time.**

**3. You can opt to take either the first or any coordinate for hotspots that are to be merged.**

**4. If the climate record has no hotspots, this means that there are no fires reported at the climate's location for that day. So, there is no need to store the cause of the fire.**

**Q4:**

**Part B Task 2**

1. It is said that a line graph is required to be plotted for air temperature against arrival time. Does “arrival time” here refers to:

a) Time in which the data visualization consumer receives the data? If this is the case, are we required to perform any additional processing to the time the consumer receives the data such as manually adding 10 seconds per data point? (i.e., arrival time = current time + 10 seconds)

b) Time in which the hotspot data is generated? (whose time is extracted directly from hotspot data)?

2. For both streaming data and static data visualization, are we strictly required to include any additional details in the visualizations besides those already mentioned, namely:

a) Maximum and minimum temperature for streaming data visualization?

b) Air temperature, surface temperature, relative humidity, and confidence for static data visualization in the map?

**Answer:**

**1. If we interpret the meaning of arrival time, this would imply that we are referring to the time at which we are consuming the message from the Kafka broker. So, you could use the current time to plot each climate record read from the Kafka broker. I do acknowledge that there could be some semantic concerns here as the current time interval may not be consistent depending on the logged climate data in the Kafka logs. In the past, we have had students applying different approaches in this context. For instance, some students may use the current time at the consumer side to represent the arrival time when reading each climate data from the Kafka broker. Other students have added a 10 second sleep interval after consuming each climate record from the broker. Therefore, you may opt to select an appropriate setting for the arrival time.**

**2. You are not required to include additional details apart from those already mentioned in Part B, Task 2.2.b in the assignment. Nevertheless, if you would like to include additional information which could improve the analysis of the results, please do so.**

**Q5:**

Are we allowed to use aggregation for Part A task 2.2?

**Answer: Yes, you are allowed to use aggregation for Part A Task 2.2.**

**Q6:**

In Part A, when questions ask for a result between two values (e.g. temperature between 65 °C and 100 °C), are those values exclusive or inclusive?

For example, let x be any temperature that lies in the range of the query results. Will it be such that:

$65 < x < 100$, OR

$65 \leq x \leq 100$?

**Answer: The question did not explicitly mention if the rage should include the upper and lower limits. Nevertheless, you can consider it as $65 \leq x \leq 100$**

**Q7:**

For Part B Task 1(d), when storing climate data in MongoDB, what should we store the station ID as? The station ID is not given in climate_streaming.csv.
Answer: You can set a constant value for the station id.

**Q8:**

1. If there are multiple climate records from the same location but with different dates (basically, the same station had multiple records), which climate record should the AQUA and TERRA hotspot records be assigned to?

2. If we receive more than 2 records for the same AQUA or TERRA producer, do we merge all of them and find the average confidence?

**Answer:**

**1. If climate producer send data and spark streaming received data every 10 second, each batch of data would have only 1 climate data. In the rare event that you receive 2 climate data in a single batch, you can handle it as outlined in the footnote of the assignment specs; that is to choose only 1 climate and drop the other.**

**2. Yes, you merge all hotspot, for example, 2 aqua and 2 terra are in the same location with similar created time, we combine 4 to 1. If we have 2 aqua and 0 terra, we combine 2 to 1.**

**Q9:**

For each question in Task A2 (PyMongo), how should we return the results?

**Answer: You should use print() or pprint()**

**Q10:**

Do we have to store the data stream generated from Part B into the same collection in Part A, or create a new one in Part B?

**Answer: You can create a separate collection for Task B, but the data model must the same to that of Task A.**

**Q11:**

Part A

1.F asks to find the number of fires for each day. Are we meant to output just the days which have hotspots and their respective number of fires OR are we meant to output the days which have no hotspots and return a value of 0 alongside the days with hotspots?

1.H asks to find the average surface temperature for each day. So in our output are we only meant to return the days with hotspots and the average temperature OR are we meant to include the other days which have zero hotspots too and just return that the average surface temperature is 0?

**Answer:**

**1.F - Please output the days which have no hotspots and return a value of 0 alongside the days with hotspots**

**1.H - Include N/A or None for the days which have zero hotspots.**

**Q12:**

Most of the time, due to a delay in starting the producers, my hotspot dates don't fall perfectly into the batch windows. So out of the 5 hotspots that are produced for each date, from each satellite, some of them will end up in the same batch as the climate data of the same date, while others won't. In this case, for each batch, do we only consider hotspot data of the same date as the climate data or do we assume that all hotspot data within a given batch is of the appropriate date and edit the date accordingly.

**Answer: All hotspot data in the same batch as the climate data will follow the climate data's date. Remember that you don’t have to append date information in the hotspot streaming data, only created time.**

**Q13:**

Task 2.2a asks us to create a bar chart detailing how many fire records there are for each hour. This implies that time data must also be stored in the MongoDB collection. However, we average across multiple records with different times (but the same date) to create a fire event (at least in my understanding). So which of the following is true,

a) We consider satellite data in the same batch but with different hours to be different fire events and don't average them into a single fire event.

b) We do combine satellite data in the same batch but with different hours into a single fire event but we also store each of the times somewhere in the database.

c) We do combine satellite data in the same batch but with different hours into a single fire event and we label this event with one of the possible times (or possibly the average).

**Answer: Consider option (a). We group all data based on hour regardless of the date. Therefore the x-axis has a range between 0 and 23 hours (or 12 am - 11pm).**

**Q14:**

In the assignment data the climate streaming data, in the precipitation column the first row the precipitation have a blank space behind it, so when we read the data using csv dictReder, it will give precipitation a blank space as well. Should we delete the blank space to make it the same as other data. Or do we alter our code to much this situation. What i have done is commenting in the code about this situation. This may occur error in the marking process depends on student's code. So that is why I point this out.

Answer: You may opt to delete the blank space from the precipitation column in the header row of the climate historic and streaming csv files. If you do this, please submit your modified csv files along with the assignment code submission. This is to ensure that your program works with your modified csv files.

**Important: Please do not modify any other aspects of the csv files which could alter the semantics of the assignment specifications, apart from what was mentioned above.**