Problem Statement for Sr. Data Engineer (Python) at Carnot

Objective: To design a service that flags when a user visits any of the hotspot locations

Success Criteria: The service will be considered successful if all hotspot visits are correctly identified along with its details for a given stream of data

Data: Refer to the attached excel for raw data and hotspot data. Assume that every entry from the excel is one data point and every data point is received in a stream at a time = time_stamp column value

Problem Statement:

Assume that you have a stream of incoming data with location details (x & y coordinates). You are also given a list of hotspot locations. It can be assumed that the hotspot location has been visited if any of the raw data point is within 5-unit radius from hotspot location.

1. What approach would you follow to check if user has visited the hotspot? WHY?
2. Where and how would you store the hotspot locations when you are making the service live for N users?
3. Using raw data excel as your stream, write the code for this service which identifies all visits to any of the hotspots and corresponding visit time. The output can be stored in another excel.
4. Time the visit processing block. Assuming the processing happens for one data point at a time, what is the average time required for your model?
5. When running this pipeline at scale, which sub-block is likely to cause the bottleneck? How would you tackle it to make overall service faster?

Note: You can make any assumptions that you might feel are necessary as long as you state them clearly in your response