For First we need to have a table to keep track of the Patients records and that must contains this attributes-

- Patient ID
- Visit Date
- Patient's Sex
- Patient's Age at the time of visit

The query is a simple one but the challenge is in the time constraint of the 200 ms

My strategy to a tackle this problem will be as follows:

1- Naive approach, I will choose a fast and efficient query language like SQL
I will select from the table where the visit date within the past 10 years and i will group by month and gender. if it fails to meet the response time requirement I will try the second approach.

2- query optimization, i will retrieve the columns that i need only (query tuning), if is still a failure, i will go for the 3rd approach

3- table splitting, i will try to split the table into smaller related ones using normalization, denormaization and partitioning , if it is still not a success i will go for the next step

4- Indexing, I will create an index on the visit date or the visit date with the sex as they are the most accessed attributes and will help query get faster, example of index if this fails , i will do the 5th approach

5- concurrency, I will choose multi-Threading techniques, I can have a thread for every month in the past 10 years. In every thread i will select the patients of this month (note i will be still using the index), benefit of this they will be executed in parallel on a different cores of the server