Rapport : Solution of DureeVisitePOI

1. Python UDF: `dureeVisite` Function

The Python function `dureeVisite` traverses through a list of `(poiID, timestamp)` pairs and calculates the duration of consecutive visits to the same POI, returning a list of `(poiID, duration)` tuples.

- **Slow and Fast Pointers**: The `slow` pointer marks the start of a particular POI, and the `fast` pointer traverses through the list. When the `fast` pointer encounters a different `poiID`, the function calculates the time difference from the `slow` pointer's start position to the previous timestamp (`fast - 1`).
- **Handling the Last Segment**: After the list is fully traversed, the function checks the final segment of consecutive POIs to compute its duration.
- **UDF Registration**: The function is registered as a Spark SQL UDF (User-Defined Function) to be used in subsequent SQL queries.

2. SQL Query Workflow

Step 1: Creating a Temporary View `DureeVisitePOI`

```
create or replace temp view DureeVisitePOI as
select
    userID, seqID,
    explode(dureeVisite(collect_list((poiID, cast(dateTaken as Integer))))) as POI_dur
ee
from user_visits
group by seqID, userID;
```

- `collect_list`: Aggregates all `poiID` and `dateTaken` values into a list for each `userID` and `seqID`.
- `dureeVisite` **UDF**: Applies the `dureeVisite` function to the aggregated list, calculating the visit durations for each POI within the sequence.
- `explode`: Unpacks the list returned by the UDF into individual rows, where each row contains a `poiID` and its corresponding `duree`.

Step 2: Calculating the Average Visit Duration and Sorting

```
create or replace temp view DureeMoyenneVisitePOI as
select
    POI_duree.poiID as poi,
    AVG(POI_duree.duree) / 60 as duree_visite_moyenne
from DureeVisitePOI
group by poi
order by duree_visite_moyenne desc;
```

- `AVG`: This calculates the average duration for each POI, converting the result from minutes to hours (assuming the input is in minutes).
- `group by`: Groups the data by `poiID` to aggregate all visit durations for each POI.
- `order by`: Sorts the results in descending order of average visit duration, so the POIs with the longest average visit times appear first.

3. Querying and Returning Results

This final query simply selects and displays all POIs along with their average visit duration, as calculated in the previous step.