Behavix Test Assignment

Sample data specification

- File name: `test_assignment_1000.parquet`
- Format: parquet
- Schema:
  - `event_time` - timestamp
  - `event_id` - string
  - `user_id` - string
  - `domain` - string
- Number of events: 1761698
- Number of days: 3
- Number of users: 1000

Note: please do not share the test sample publicly.

Task

Your task is to write a script that calculates web sessions based on the sample data.

A session is a sequence of events from the same user identified by unique session_id. A new session begins if there is a gap of more than 30 minutes between events.

For each session, you need to calculate:

- session_id: A unique identifier for each session.
- event_duration: The time difference between consecutive events within the session. For the last event in a session, this should be NULL.
- domain_duration: The total time spent on each domain during the session.

It should be possible to join the resulting session table back to the original sample using `event_id`. As a result of such join, all events that belong to the same session should get the same `session_id` and a correct `event_duration`.

You may use any programming language and toolset of your choice. However, the solution must not rely on any cloud services such as AWS Athena or Glue, it should run locally. Local Spark is ok to use.

The result is stored as a parquet file:

```
calculate_web_sessions test_assignment_1000.parquet -o sessions.parquet
```

Delivery result

The result of this test assignment should be delivered as an archive of source files or a git repository. It should include documentation on how to set up, build and run the code.

Think about possible improvements to the schema and the solution, share your thoughts.