User routes on the site

Description

Clickstream is a sequence of user actions on a website. It allows you to understand how users interact with the site. In this task, you need to find the most frequent custom routes.

Input data

Input data is a table with clickstream data in file `hdfs:/data/lsml/sga/clickstream.csv`.

Table structure

- `user_id (int)` - Unique user identifier.
- `session_id (int)`- Unique identifier for the user session. The user's session lasts until the identifier changes.
- `event_type (string)` - Event type from the list: **page** (visit to the page), **event** (any action on the page), **<custom>** (string with any other type).
- `event_type (string)` - Page on the site.
- `timestamp (int)` - Unix-timestamp of action.

Browser errors

Errors can sometimes occur in the user's browser - after such an error appears, we can no longer trust the data of this session and all the following lines after the error or at the same time with it are considered corrupted and **should not be counted** in statistics.

When an error occurs on the page, a random string containing the word **error** will be written to the `event_type` field.

Sample of user session

```
+-------+----------+------------+----------+----------+
|user_id|session_id|  event_type|event_page| timestamp|
+-------+----------+------------+----------+----------+
|    562|       507|        page|      main|1620494781|
|    562|       507|       event|      main|1620494788|
|    562|       507|       event|      main|1620494798|
|    562|       507|        page|    family|1620494820|
|    562|       507|       event|    family|1620494828|
|    562|       507|        page|      main|1620494848|
|    562|       507|wNaxLlerrorU|      main|1620494865|
|    562|       507|       event|      main|1620494873|
|    562|       507|        page|      news|1620494875|
|    562|       507|        page|   tariffs|1620494876|
|    562|       507|       event|   tariffs|1620494884|
|    562|       514|        page|      main|1620728918|
|    562|       514|       event|      main|1620729174|
|    562|       514|        page|   archive|1620729674|
|    562|       514|        page|     bonus|1620729797|
|    562|       514|        page|   tariffs|1620731090|
|    562|       514|       event|   tariffs|1620731187|
+-------+----------+------------+----------+----------+
```

Correct user routes for a given user:

- **Session 507**: main-family-main
- **Session 514**: main-archive-bonus-tariffs

Route elements are ordered by the time they appear in the clickstream, from earliest to latest.

The route must be accounted for completely before the end of the session or an error in the session.

Problem

You need to use the Hive, Spark RDD and Spark DF interfaces to create a solution file, the lines of which contain **the 30 most frequent user routes** on the site.

Each line of the file should contain the `route` and `count` values **separated by tabs**, where:

- `route` - route on the site, consisting of pages separated by "-".
- `count` - the number of user sessions in which this route was.

The lines must be **ordered in descending order** of the `count` field.

How to submit

When you're ready to submit, you can upload files for each part of the assignment on the "My submissions" tab.