Data Engineer task

Context
You own a website and you're selling advertisement space to monetize it.

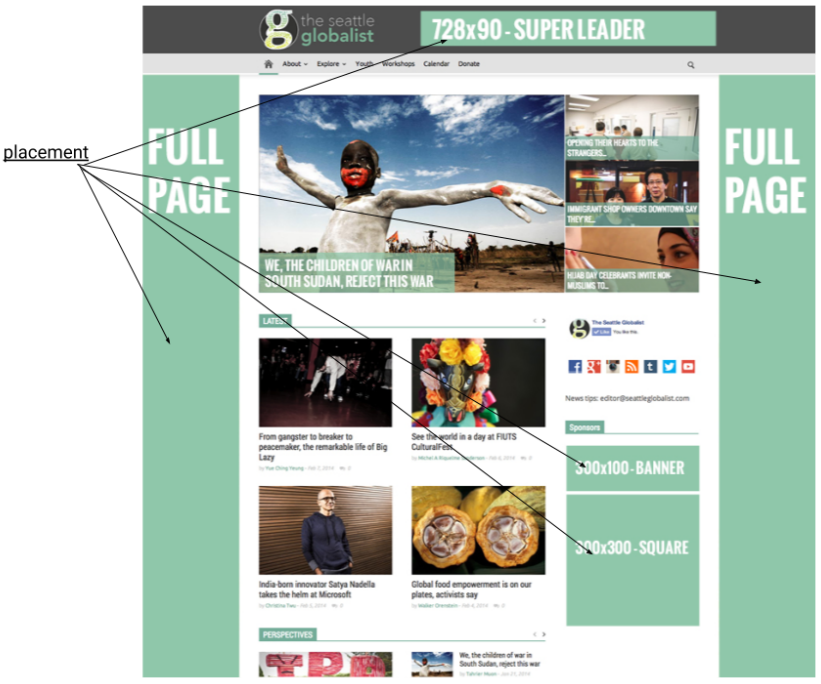
For every single banner view or click, you are receiving an event.
You may assume that events are received hourly in batch files or, if you prefer, in real time in streaming - whichever you feel most comfortable with.

For every single view or click, you'll get an event containing the following data:

- Timestamp: long, unix timestamp - UTC timezone
- Event_type, integer, [0=impression, 1=click]
- Banner_id, string
- Placement_id, string
- Page_id, string
- User_id, string

Where

- user_id is the cookie id of the user who clicked or viewed a specific ad. If the same user sees or clicks multiple ads, he will have the same user_id.
- Banner_id is the id of the specific "image" shown on the website.
- Page_id is the id of the page (that is, an id corresponds to a URL)
- Placements are locations where ads can appear on a website. Each placement has a unique placement_id. See the image below as an example.



Task

- Write a simple script to generate a dataset containing a sample of above data.
- Write a spark job to compute hourly and daily statistics for each placement_id (number of views and clicks of each placement_id).
- Update the above job to include the count of distinct users which viewed or clicked on a placement.
- Create a new spark job to compute the same statistics above (number of views, clicks and distinct users) for each page.

Please assume that the number of events is quite large, as this is a very busy website!

Spark code can be written in pyspark, scala or any other language supported by spark. Feel free to use a different framework other than spark if you prefer, as long as it is opensource.

Deliverables

- Provide a git repository or an archive file containing all the source code you've written.
- Provide all the instructions to run the written jobs.
- Optionally, the jobs should run using docker.
- Provide a README file.