

Brochure page views exercise

There are five JSON files:

brochures.json - list of brochure objects

brochure_clicks.json - list of the brochure click events

page_turns.json - list of the page turn events

enters.json - list of the page enter events

exits.json - list of page exit events

Brochure object contains information about brochure, but it's not relevant for this exercise at least in this level.

Brochure click event is generated when the user opens a brochure in app or web browser. It contains:

- *date_time* – timestamp when the brochure was clicked
- *brochure_click_uuid* - a unique identifier of this event
- *brochure_id* - ID of the brochure that was opened
- other user session information

Enter event is tracked at the same time when brochure click event is generated. Basically, it says what page of a brochure user opens first, it contains:

- *brochure_click_uuid* - a unique identifier of this event, which links it to the brochure click event
- *date_time* - timestamp when the page was entered
- *page* - number of a page user opens first (user can be landed on every page of a brochure, but most of the time it is page 1 or 0 which has to be treated as 1)
- *page_view_mode* - mode of brochure viewer. It is either SINGLE_PAGE_MODE or DOUBLE_PAGE_MODE (portrait or landscape mode). In the portrait mode user sees only a single page, in the landscape mode user sees two pages (think about a book or magazine, cover is page 1)

Exit event is tracked when the user leaves the brochure (closes brochure viewer). It contains:

- *brochure_click_uuid* - which links it to the brochure click event
- *date_time* - timestamp when the exit event
- *page* - last page number that was opened before the user closed the brochure
- *page_view_mode* - mode of brochure viewer. It is either SINGLE_PAGE_MODE or DOUBLE_PAGE_MODE (portrait or landscape mode). In the portrait mode user sees only a single page, in the landscape mode user sees two pages (think about a book or magazine, cover is page 1)

Page turn event is tracked when the user turns the page in the brochure viewer. This event data contains:

- *brochure_click_uuid* - which links it to the brochure click event
- *date_time* - timestamp when the page was turned
- *page* - number of page user turns to (if the user was on page 3 and turns to page 4, then the page value is 4)
- *page_view_mode* - mode of brochure viewer. It is either SINGLE_PAGE_MODE or DOUBLE_PAGE_MODE (portrait or landscape mode). In the portrait mode user sees only a single page, in the landscape mode user sees two pages (think about a book or magazine, cover is page 1)

Task is to write a Spark job that creates a dataset with user activities objects (JSON format). User activities object is supposed to be an information about the total number of pages the user entered, the total number of pages user exists, and the total number of pages the user viewed. So, it must have the following fields:

- *User_ident* - The user identifier
- *Total_enters* - Total number of pages the user entered
- *Total_exits* - Total number of pages the user existed
- *Total_viewed* - Total number of pages the user viewed