Spark Assignment

Write a spark application to transform ad events data

Ad Events Data

You start with data consisting of ad event per line. An ad event is represented by event id, timestamp, type, visitorId, and pageUrl. Each user will have a unique visitorId associated to it. Dataset can be downloaded from:
https://s3.amazonaws.com/de-coding-challenge/ad-events-2018060100.tar.gz

Data Transformation

You need to write a Apache Spark application to analyze ad event to find user journey through each webpage.

Each line represents an ad event and each user will have unique “visitorId”. You are required to find “nextPageUrl” user visited per event during their ad journey. The output file should have 6 columns: id, timestamp, type, visitorId, pageUrl and nextPageUrl.

Source Input Schema:

id: string, timestamp: string, type: string, visitorId: string, pageUrl: string

Expected Output Schema:

id: string, timestamp: string, type: string, visitorId: string, pageUrl: string, nextPageUrl: String

For example, given two sequential in time events e 1 and e 2 from a visitor v1 , the output event o1 should contain the id, timestamp, type, visitorId, & pageUrl from e1 but the nextPageUrl will be the pageUrl from e2 .

Solution Requirements:

- Must use functional api either with Dataset, Dataframe or RDD
- Must work on bigger source data with millions of events per visitor during a given hour
- Spark SQL raw queries will be not acceptable