Click Stream Web Analytics

**Purpose:**

Generate various analytics for web users through click stream web data. On a web site, click stream analysis is the process of collecting, analysing, and reporting aggregate data about which pages visitors visit in what order as a result of succession of mouse click each visitors makes

**Inputs:**

The input would be the click stream data generated byjava tool in the JSON format. Number of records should be at least a million.

**Steps:**

* + The data can be generated through a Java program that generates data in JSON format in a file.
  + The above raw data would be stored in HDFS and then partitioned on the basis of any appropriate data field for example: date/timestamp, page, user, location etc. This data loading can be done using MR Job.
  + Create a mechanism/process to load data through daily incrementaI/ingestion process. This can be a sequence of MRjobs. Schedule this incrementaI/ingestion process through any appropriate mechanism (Cron job, Oozie work flow, Spring batch).
  + Create appropriate external tables in Hive preferably using MySQL Meta store.
  + The above tables should be partitioned and bucketed on appropriate data fields.
  + Perform various analytic on the above data using Hive, Pig and store the analytical results in any NoSQL data store (HBase).
  + Visualise the analytical results through any reporting tool (Jasper, Tableau etc).

**Analytics:**

* Generate following analytics/reports using above data:
* Traffic analytics
* Most visited pages by users.
* Location wise traffic analysis
* How many times back or refresh button used.
* Perform other appropriate analytics based on the traffic data.

**Ecommerce Analytics**

* User item visit analysis.
* Users purchase/shopping cart analysis.
* Perform other appropriate ecommerce analytics base on the data.

**Deliverables**

* + Approach document: end of week # 1
  + Case study presentation: end of week # 4
  + Code Artifacts: end of week # 4
  + Output report: end of week # 4

**Schema**

The data generation should be confined with the below schema

* + userID – Unique id of the user
  + url -- to keep track of user activity
  + action (optional)-- to keep track of action performed while performing the cart activity

Can be covered as part of url or

Possible actions : view , addtocart , removefromcart , purchase etc (Back and refresh are not part of this)

* + location – to keep track of the location of the user
  + logTime – to keep track of the time when the user logged in
  + payment\_method – to keep track of the payment method opted by user
  + logDate – to keep track of the date when the user logged in

For e.g.

{"userID":"20","url":"http:\/\/www.shop.com\/homePage", "action":"view","location":"Boston","logTime":119852762544685,"payment\_method":"CreditCard",logDate":"07-21-2017"}

{"userID":"4","url":"http:\/\/www.shop.com\/homePage", "action":"addtocart","location":"Memphis","logTime":119852762545005,"payment\_method":"NetBanking","logDate":"07-21-2017"}