**Data Quality Patterns Lab**

**User Growth Pipeline**

We want to measure the website traffic and user growth for a startup that sells bootcamps and courses.

The goal of this pipeline is to answer the following questions:

* How many people are going to the website on daily basis?
  + What is the geographical and device break down of that traffic?
  + Where are these people coming from? Linkedin? Substack?
* How many people are signing up with an account on the website each day?
  + What percentage of traffic is converting to signing up?
* How many people are purchasing the boot camps and courses of that website?
  + What percentage of signups convert to paying customers?

**Business Metrics**

|  |  |  |
| --- | --- | --- |
| **Metric Name** | **Definition** | **Is Guardrail** |
| **Signup\_converstion\_rate** | COUNT (signup)/ COUNT (website\_hits) | YES (problem with the landing page) |
| **Purchase\_conversion\_rate** | COUNT (purchases) / COUNT (signups) | YES (problem with the checkout page) |
| **Traffic\_breakdown** | COUNT (website\_hits) GROUP BY referrer (linkedin, insta, fb) | NO (AB test is not going have no impact on where the traffic comes from) |

Guardrail metrics signifies a problem to the business that is really bad. It is protecting the business. Especially in the case of an AB test.

**Flow Diagram**

A diagram of a diagram

Description automatically generated

**Schema**

**Core.fact\_website\_events** : This table is a list of all events for the website and includes IP enrichment and user enrichment for country and device specific information.

The unique identifier for this table is logged\_out\_user\_id and event\_timestamp.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Column Type** | **Column Comment** |
| User\_id | BIGINT | This column is nullable for logged out events. This column indicates the user who generated this event. |
| Logged\_out\_user\_id | BIGINT | This column is a hash of IP address and device information. (we’ll use the murmur3 hash) |
| Dim\_hostname | STRING | The host associated with this event (website url). |
| Dim\_country | STRING | The country associated with the IP address of the request. |
| Dim\_device\_brand | STRING | The device brand associated with this request. Could be NULL because of bots that don’t have a brand. |
| Dim\_action\_type | STRING | This is an enumerated list of actions that a user could take on this website (signup, watch video, go to the landing page etc) |
| Event\_timestamp | TIMESTAMP | The UTC timestamp for when this event occurred. |
| Other\_properties | MAP[String, String] | Any other valid properties that are part of this request. |
| Ds | STRING | This is the partition column for this table |

**Quality Checks**

* Not null checks on (dim\_hostname, dim\_action\_type, event\_timestamp, logged\_out\_user\_id, dim\_country)
* No duplicates on Primary Key
* Dim\_hostname is well\_formatted (something like [www.x.com](http://www.x.com))
* Row count checks
  + GROUP ON dim\_hostname and check week-over-week counts of the website hosts.
* Enumeration checks on dim\_action\_type (should be signup, purchase, login, etc)

**Core.agg\_website\_events:** This table is an aggregated view of all website events.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Column Type** | **Column Comment** |
| Dim\_action\_type | STRING | This is an enumerated list of actions that a user could take on this website (signup, watch video, go to the landing page etc) |
| Dim\_country | STRING | The country associated with the IP address of the request. |
| Dim\_device\_brand | STRING | The device brand associated with this request. Could be NULL because of bots that don’t have a brand. |
| Event\_hour | INTEGER | The hour this event took place in UTC. |
| M\_total\_events | BIGINT | The total number of events for this slice |
| Aggregation\_level | STRING | This is how this aggregate table is GROUPED. Values include (dim\_country\_dim\_action\_type\_dim\_device\_brand\_event\_hour, dim\_action\_type, dim\_country\_dim\_action\_type, (overall)) |
| Ds | STRING | This is the partition column for this table. |

**Quality Checks**

* Row count checks
  + (overall) rollup should have more data than any other rollup
* Event\_hour should look like it’s old seasonal pattern
* M\_total\_events should be > some minimum number