**Challenge 1:**

* I would create an order-product mapping table:  
  `ae\_data\_challenge\_v1\_akunapuli.order\_product\_mapping`

with orders as (

select line.product\_id as product\_id, o.\_id as order\_id, o.created\_at

from `ae\_data\_challenge\_v1.orders` o, UNNEST(line\_items) as line)

, products as(

select var.sku as product\_sku, var.option1 as product\_style, var.option2 as product\_size, p.category, p.title, p.\_id

from `ae\_data\_challenge\_v1.products`p, UNNEST(variants) as var

)

select o.\*, p.\* from orders o join products p

on o.product\_id = p.\_id

* New Web Events created table:

- - Defining a session  
WITH cte AS (

SELECT

cookie\_id,

timestamp,

event\_url,

utm\_medium,

utm\_campaign,  
 utm\_source

lag(timestamp) OVER (PARTITION BY cookie\_id ORDER BY timestamp) AS prev\_event\_time,

row\_number() OVER (PARTITION BY cookie\_id ORDER BY timestamp) AS session\_id

FROM `ae\_data\_challenge\_v1.web\_events`

),

(\*\*\*defining a session \*\*\*)

sessions as (SELECT

cookie\_id,

session\_id,

count(\*) AS session\_count,

timestamp

FROM cte

WHERE

(timestamp - prev\_event\_time) <= INTERVAL '30' MINUTE

OR prev\_event\_time IS NULL

GROUP BY

cookie\_id,

session\_id) ,

min\_session AS (

SELECT

session\_id,

MIN(timestamp) AS start\_timestamp

FROM sessions

GROUP BY session\_id

) ,

(\*\*\*start session timestamp \*\*\*)

start\_session\_timestamp as (SELECT

sessions.session\_id,

min\_session.start\_timestamp,

sessions.timestamp

FROM sessions

JOIN min\_sessions

ON sessions.session\_id = min\_sessions.session\_id

AND sessions.timestamp = min\_sessions.start\_timestamp) ,

(\*\*\*first event URL of session \*\*)

session\_event\_URL as (SELECT

m.session\_id,

a.event\_url,

m.start\_timestamp

FROM cte a

JOIN min\_sessions m

ON cte.session\_id = min\_sessions.session\_id

AND cte.timestamp = min\_sessions.start\_timestamp) ,

(\*\*\*first UTM Medium of session \*\*\*)

Utm\_medium as (SELECT

m.session\_id,

a.utm\_medium,

m.start\_timestamp

FROM cte a

JOIN min\_sessions m

ON cte.session\_id = min\_sessions.session\_id

AND cte.timestamp = min\_sessions.start\_timestamp) ,

(\*\*\*first UTM Source of session \*\*\*)

Utm\_source as (SELECT

m.session\_id,

a.utm\_source,

m.start\_timestamp

FROM cte a

JOIN min\_sessions m

ON cte.session\_id = min\_sessions.session\_id

AND cte.timestamp = min\_sessions.start\_timestamp) ,

(\*\*\*first UTM Campaign of session \*\*\*)

Utm\_campaign as (SELECT

m.session\_id,

a.utm\_campaign,

m.start\_timestamp

FROM cte a

JOIN min\_sessions m

ON cte.session\_id = min\_sessions.session\_id

AND cte.timestamp = min\_sessions.start\_timestamp)

Select a.timestamp, a.event\_url, a.utm\_medium, a.utm\_campaign, a.utm\_source, b.session\_id, c.\*, d.\*, e.\*, f.\*, g\* from cte a, sessions b, start\_session\_timestamp c, session\_event\_URL d, Utm\_medium e , Utm\_source f , Utm\_campaign g

**Challenge 3:**

Some Data Quality Issues in Orders table:

* \_id is the primary key in the orders table, but a lot of ids have duplicates. The highest being 18 for id: B1802901
* There are a lot of order ids that have 0 subtotal but some amount $ of total- this could make sense if the customer had discounts and got the product for free but only had to pay for shipping etc.
* There are 2387 ids that didn’t have to pay anything (i.e. total = 0, subtotal=0).
* Loaded\_at is always greater than created\_at. But the time taken to load after created is varying from id to id. The least time was 1 hour 12 min and the max time was approximately 134 days.

Some Data Quality issues in Products table:

* Inconsistencies found. For some combinations of Product Id and Variant Id, there were more than one values of Updated time of variant across entries.
* The Variables option1 field has a lot of inconsistent data. For example: it has $50, $100, but also has King size, Queen size field entries.
* Variable option2 has a lot of NULL entries making it inconsistent.
* \_id is not unique and there’s duplicates. Highest being 38577
* Variants\_title has inconsistent data. There’s some fields that describe the size as king size, twin etc and others that have measurements like 9 \* 12.

Some Data Quality issues in Web Events:

* Unique identifier for events- \_id has multiple duplicates.
* There’s 1478750 NaN values in the customer\_id field making it inconsistent.
* There are 1270791 NULLs in the customer field.
* For the event\_name email sign up, the URL is missing.