About the data

There are three source tables loaded into our data warehouse (Google BigQuery) containing a month's worth of data collected. Please access BigQuery at this link to view and query source data (you have READ access) and your target BigQuery dataset (you have WRITE access).

orders

Description:
Contains a record for each order placed through our e-commerce system
Schema:

COLUMN NAME	DATA TYPE	DESCRIPTION
_id	STRING	Unique identifier for an order
_loaded_at	TIMESTAMP	Timestamp to indicate when the record was loaded into the table
created_at	TIMESTAMP	Timestamp to indicate when the order was created
updated_at	TIMESTAMP	Timestamp to indicate when the order was last updated
subtotal	NUMERIC	The dollar sum of all line item amounts after discounts but before shipping, taxes, and tips (in USD)
total	NUMERIC	The dollar sum of all line item amounts, discounts, shipping, taxes, and tips (in USD)
line_items	RECORD [REPEATED]	Array of line item objects purchased in the order

products

Description:
Contains a record for each product available for purchase in our stores
Schema:

COLUMN NAME	DATA TYPE	DESCRIPTION
_id	INTEGER	Unique identifier for a product
_loaded_at	TIMESTAMP	Timestamp to indicate when the record was loaded into the table

category	STRING	Product category used to group products together
created_at	TIMESTAMP	Timestamp to indicate when the product was created in our e-commerce system
updated_at	TIMESTAMP	Timestamp to indicate when the product was last updated in our e-commerce system
title	STRING	Title of the product
variants	RECORD [REPEATED]	Array of all existing product variants associated with the product, each variant representing a different version of the product.

web_events

Description:
Contains a record for each event committed by a user on our web store
Schema:

COLUMN NAME	DATA TYPE	DESCRIPTION
_id	STRING	Unique identifier for the event
_loaded_at	TIMESTAMP	Timestamp to indicate when the record was loaded into the table
cookie_id	STRING	Device identifier used to indicate a website visitor. For a new visitor, this value is set in the user's browser cookies.
customer_id	STRING	Unique identifier to indicate a customer. This value is null if the user is an anonymous web visitor
event_name	STRING	Event name that indicates how user interacted with the website
event_url	STRING	URL on which the event occurred
event_properties	STRING	JSON string containing contextual properties relating to the event
timestamp	TIMESTAMP	Timestamp to indicate when the event occurred
utm_campaign	STRING	Campaign that referred the user
utm_medium	STRING	Medium that referred the user
utm_source	STRING	Source that referred the user

Challenge #1 - data modeling

Boll & Branch business analysts need reporting tables that they can query using a business intelligence tool. Based on the requirements below, please create a new dbt project that transforms the raw source data into reporting tables using dbt-SQL. In addition to your dbt code, please include:

1. An entity relationship diagram that shows the relationship between each of the tables you intend to expose to the analysts. Include relevant fields and the joinable keys. You can use pencil and paper, README, or any digital drawing or diagramming tool.
2. Documentation of each of your metrics' specifications (i.e. how a BI tool should define them in SQL). You can use dbt Metrics or just include an extra file

Once your code, diagram, and docs are ready for review, check them into a shareable git repository.

The analysts' requirements:

- Need the ability to report on the following metrics:
    - ***Total Order Count***: the count of orders
    - ***Total Gross Revenue***: sum of total line item revenue minus sum of line item discounts
    - ***Total Order Units***: sum of order line item quantities
    - ***Average Order Value***: average of order subtotals
    - ***Average Order Units***: average of units per order
    - ***Total Pageviews***: count of `page` web events
    - ***Total Web Sessions***: count of web sessions, where a 'session' is defined as a series of one or more web events committed by the same cookie with no more than a 30 minute gap between events. Any 30 minute gap indicates a new session.
    - ***Total Bounced Web Sessions***: count of web sessions where total `page` events is less than or equal to 1
    - ***Total Web Users:*** distinct count of users in web sessions, where a user is defined as the first known customer_id that is associated with the cookie of the web session. If the session user has no known customer_id then default to the cookie_id.
    - ***Bounce Rate***: Total Bounced Web Sessions divided by Total Web Sessions
    - ***Product View Rate***: Total web sessions that include a `product_viewed` event divided by Total Web Sessions
    - ***Add To Cart Rate***: Total web sessions that include a `product_added` event divided by Total Web Sessions
    - ***Checkout Rate***: Total web sessions that include a `checkout_step_viewed` event divided by Total Web Sessions
    - ***Signup Rate***: Total web sessions that include a `email_sign_up` event divided by Total Web Sessions

    - ***Conversion Rate***: Total web sessions that include a `order_completed` event divided by Total Web Sessions
- Need the ability to filter and breakdown total order count, units, and revenue by dimensions:
    - ***Order Created Timestamp***
    - ***Product Category***
    - ***Product Title***
    - ***Product SKU*** - based on the options available in product variants
    - ***Product Style*** - based on the options available in product variants
    - ***Product Size*** - based on the options available in product variants
- Need the ability to filter and breakdown total web sessions, pageviews, and rates by dimensions:
    - ***Web Session Start Timestamp*** - timestamp of the first event of a session
    - ***Landing Page URL*** - first event URL of a session
    - ***Session Medium*** - first `utm_medium` of a session
    - ***Session Source*** - first `utm_source` of a session
    - ***Session Campaign*** - first `utm_campaign` of a session

Challenge #2 - ad hoc queries

Answer the following questions from Boll & Branch analysts using the source data samples in the data warehouse. Please commit your answers and SQL queries in a folder named `analysis` in your git repository.

1. What proportion of orders contained a product with category “Sheet Sets” and size “King”?
2. Which product SKU generated the most gross revenue?
3. Which date had the highest average order units?
4. What was the conversion rate of web sessions where the user added a “Plush Bath Towel Set” product to their cart?
5. Among `page` events only, what are the top five most common page URLs that immediately preceded a user’s navigation to the “checkout.bollandbranch.com” domain during their session?
6. What are the top five non-null session campaigns which garnered the most web users?
7. What are the top five non-null session sources which garnered the most gross revenue?

Challenge #3 - data quality review

Using the programming language of your choice (SQL, Python, R, Bash, etc...) identify any data quality issues you came across while working with the source data. We are not expecting a full blown review of all the data provided, but instead want to know how you explore and evaluate

data of questionable provenance. Please commit any exploratory code and your findings to the git repository.