PART 1 Basic Analytics using BigQuery

P1) Count number of Rows all sessions raw

SELECT

  COUNT(\*) as count

  FROM `data-to-insights.ecommerce.all\_sessions\_raw`

P2) Count number of rows all sessions

SELECT

  COUNT(\*) as count

  FROM `data-to-insights.ecommerce.all\_sessions`

P3) Check min and max date of all sessions

SELECT

  MIN(PARSE\_DATE('%Y%m%d', CAST(date AS STRING))) as MinDate,

  MAX(PARSE\_DATE('%Y%m%d', CAST(date AS STRING))) as MaxDate

 FROM `data-to-insights.ecommerce.all\_sessions`

P4) Query to show total unique visitors from both all sessions and all sessions raw datasets

SELECT

  count(pageviews) as total\_pageviews,

  count(DISTINCT fullVisitorID) as total\_unique\_visitors

  FROM `data-to-insights.ecommerce.all\_sessions`

SELECT

  count(pageviews) as total\_pageviews,

  count(DISTINCT fullVisitorID) as total\_unique\_visitors

  FROM `data-to-insights.ecommerce.all\_sessions\_raw`

P5) Query to show total unique visitors by the referring site

SELECT

 channelGrouping,

 count(DISTINCT fullVisitorID) as total\_unique\_visitors

 FROM `data-to-insights.ecommerce.all\_sessions`

GROUP BY channelGrouping

ORDER BY total\_unique\_visitors DESC

SELECT

 channelGrouping,

 count(DISTINCT fullVisitorID) as total\_unique\_visitors

 FROM `data-to-insights.ecommerce.all\_sessions\_raw`

GROUP BY channelGrouping

ORDER BY total\_unique\_visitors DESC

P6) Query to list all the unique product names in alphabetical order

SELECT

  DISTINCT v2ProductName as products

  FROM `data-to-insights.ecommerce.all\_sessions`

ORDER BY products

SELECT

  DISTINCT v2ProductName as products

  FROM `data-to-insights.ecommerce.all\_sessions\_raw`

ORDER BY products

PART 2 Duplicate Detection in Data

P7) returns the number of duplicate records in all\_sessions\_raw table

SELECT

  COUNT(\*) as num\_duplicate\_rows,

  \*

FROM `data-to-insights.ecommerce.all\_sessions\_raw`

GROUP BY fullVisitorId, channelGrouping, time, country, city, totalTransactionRevenue, transactions, timeOnSite, pageviews,sessionQualityDim, date, visitId,type, productRefundAmount, productQuantity, productPrice, productRevenue, productSKU, v2ProductName, v2ProductCategory, productVariant,currencyCode,itemQuantity, itemRevenue, transactionRevenue, transactionId, pageTitle, searchKeyword,pagePathLevel1, eCommerceAction\_type, eCommerceAction\_step, eCommerceAction\_option

SELECT

  (SELECT COUNT(1) from (SELECT DISTINCT \* FROM `data-to-insights.ecommerce.all\_sessions\_raw`)) as distinct\_rows,

  (SELECT COUNT(1) FROM `data-to-insights.ecommerce.all\_sessions\_raw`) as total\_rows

P8) shows no duplicate records in all sessions table

SELECT

  DISTINCT \*

FROM `data-to-insights.ecommerce.all\_sessions`

GROUP BY fullVisitorId, channelGrouping, time, country, city, totalTransactionRevenue, transactions, timeOnSite, pageviews,sessionQualityDim, date, visitId,type, productRefundAmount, productQuantity, productPrice, productRevenue, productSKU, v2ProductName, v2ProductCategory, productVariant,currencyCode,itemQuantity, itemRevenue, transactionRevenue, transactionId, pageTitle, searchKeyword,pagePathLevel1, eCommerceAction\_type, eCommerceAction\_step, eCommerceAction\_option

SELECT

  (SELECT COUNT(1) from (SELECT DISTINCT \* FROM `data-to-insights.ecommerce.all\_sessions`)) as distinct\_rows,

  (SELECT COUNT(1) FROM `data-to-insights.ecommerce.all\_sessions`) as total\_rows

PART 3 Advanced analytics using BigQuery

P9) Write a query to list the five products with most views from all visitors. Count the number of times a product was viewed, descending order and top 5 products

SELECT

  v2ProductName,

  SUM(pageviews) as total\_pageviews

  FROM `data-to-insights.ecommerce.all\_sessions`

GROUP BY v2ProductName

ORDER BY total\_pageviews DESC

LIMIT 5

P10) Query that does not double-count product views who have views the product many times

SELECT

  v2ProductName,

  COUNT(pageviews) as total\_pageviews

  FROM `data-to-insights.ecommerce.all\_sessions`

GROUP BY v2ProductName

ORDER BY total\_pageviews DESC

LIMIT 5

P11) include total number of units ordered

SELECT

  v2ProductName,

  COUNT(pageviews) as total\_pageviews,

  SUM(productQuantity) as total\_units\_ordered

  FROM `data-to-insights.ecommerce.all\_sessions`

GROUP BY v2ProductName

ORDER BY total\_pageviews DESC

LIMIT 5

P12) Include average amount of products per order

SELECT

  v2ProductName,

  COUNT(pageviews) as total\_pageviews,

  SUM(productQuantity) as total\_units\_ordered,

  SUM(productQuantity)/COUNT(productQuantity) as average\_amount\_per\_order

  FROM `data-to-insights.ecommerce.all\_sessions`

GROUP BY v2ProductName

ORDER BY total\_pageviews DESC

LIMIT 5

PART 4 Schedule your analytics queries in ETL pipeline

P13) build a pipeline that runs queries you wrote for P4, P5 and P12

P14) Order P4, then P5, and P12

P14) schedule it every day

gsutil cp /Users/aidan/PycharmProjects/airflow\_dags/python\_operator\_dag.py gs://asia-east2-avkassignment3-30b2363b-bucket/dags

gsutil cp /Users/aidan/PycharmProjects/airflow\_dags/bigquery\_operator\_dag.py gs://asia-east2-avkassignment3-30b2363b-bucket/dags

Part 6 Exporting the Results

P19) Using command line operation, export the result of P12 to google cloud storage

bq extract --destination\_format=CSV --field\_delimiter=tab --print\_header=false aidanvanklaverenassignment3:Ecommerce.total\_visitors\_and\_product\_info gs://asia-east2-avkassignment3-30b2363b-bucket/data

P20) add that step as last step of pipeline

Youtube link - https://www.youtube.com/watch?v=Kt0Gp\_YTImQ&list=PL7ZQpq2JWci9bWLnr6BzliP8e8PxbPxb6&index=2