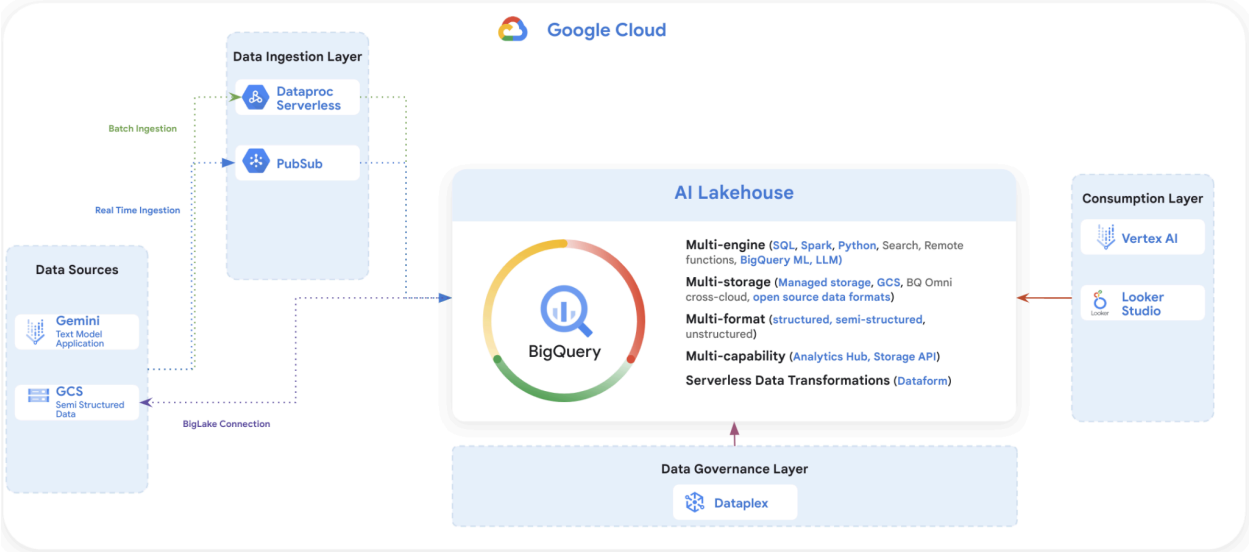
Data & AI Boot-Kon Event

Title: Data Sharing with Analytics Hub

Author: Wissem Khlifi **Date**: 2024-04-01 **Estimated Completion Time**: 45 Minutes

CAUTION:
This lab is for educational purposes only and should be used with caution in production environments. Google Cloud Platform (GCP) products are changing frequently, and screenshots and instructions might become inaccurate over time. Always refer to the latest GCP documentation for the most up-to-date information.

Architecture Diagram:



Goal of the Lab

- We have previously created the fraud detection model predictions.
- After running the dataplex data discovery job , we noticed a new BigQuery dataset created called ***"bootkon_raw_zone"*** , ***data_prediction*** biglake table were automatically created by Dataplex discovery jobs.
- The goal of the ***FraudFix*** data scientist team is to share the results of the data prediction with the customer.
- The customer will use the PCA data and perform reversed PCA in order to get the result of the predictions.
- The customer will also use the explainability results from the ***data_prediction*** data to understand why the decisions have been made to flag a given transaction as fraudulent or not.
- There is a small caveat, the ***data_prediction*** biglake table has the email address of the service account or user who has performed the machine learning tasks. This information is considered PII data (Personal Identifiable Information). In addition you should not share the auto generated transaction id that you have added during the machine learning labs.

- Data clean room within the Data Analytics Hub will allow you to share the data securely and without leaking any PII information.
- You have been assigned a group of work.
- Each group member will play the role of a data provider and a data subscriber of the other group members.
- You are a data publisher and data subscriber. You are publishing the results of your data prediction and you are subscribing to other team member data prediction.
- Collect the GCP account addresses of the your group members assigned to you, in order to set up privileges and share the data with them.

READING Section : What Analytics Hub is and what problems it solves

What is a Data Clean Room ?

Data clean rooms provide a security-enhanced environment in which multiple parties can share, join, and analyze their data assets without moving or revealing the underlying data.

BigQuery data clean rooms are built on the Analytics Hub platform. While standard Analytics Hub data exchanges provide a way to share data across organizational boundaries at scale, data clean rooms help you address sensitive and protected data-sharing use cases. Data clean rooms provide additional security controls to help protect the underlying data and enforce analysis rules that the data owner defines.

The following are primary use cases:

- **Campaign planning and audience insights.** Let two parties (such as sellers and buyers) mix first-party data and improve data enrichment in a privacy-centric way.
- **Measurement and attribution.** Match customer and media performance data to better understand the effectiveness of marketing efforts and make more informed business decisions.
- **Activation.** Combine customer data with data from other parties to enrich understanding of customers, enabling improved segmentation capabilities and more effective media activation.

There are also several data clean room use cases beyond the marketing industry:

- **Retail and consumer packaged goods (CPG).** Optimize marketing and promotional activities by combining point-of-sale data from retailers and marketing data from CPG companies.
- **Financial services.** Improve fraud detection by combining sensitive data from other financial and government agencies. Build credit risk scoring by aggregating customer data across multiple banks.
- **Healthcare.** Share data between doctors and pharmaceutical researchers to learn how patients are reacting to treatments.
- **Supply chain, logistics, and transportation.** Combine data from suppliers and marketers to get a complete picture of how products perform throughout their lifecycle.

Roles

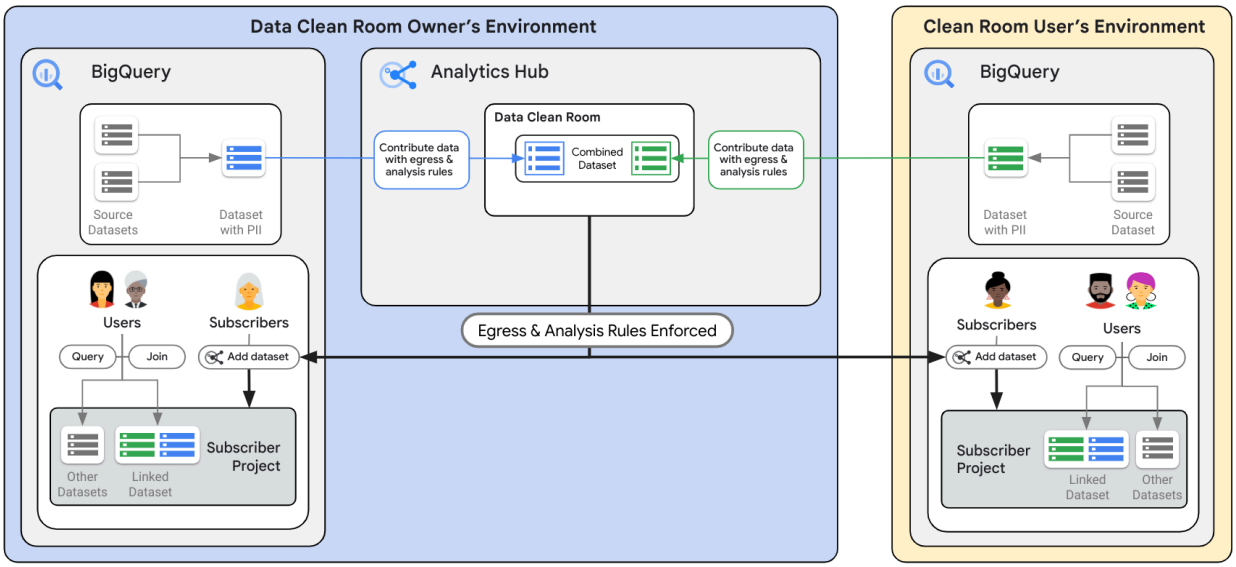
There are three main roles in BigQuery data clean rooms:

- Data clean room owner: a user that manages permissions, visibility, and membership of one or more data clean rooms within a project. This role is analogous to the Analytics Hub Admin.
- Data contributor: a user that is assigned by the data clean room owner to publish data to a data clean room. In many cases, a data clean room owner is also a data contributor. This role is analogous to the Analytics Hub Publisher.

- Subscriber: a user that is assigned by the data clean room owner to subscribe to the data published in a data clean room, letting them run queries on the data. This role is analogous to a combination of the Analytics Hub Subscriber and Analytics Hub Subscription Owner. Subscribers must have non-edition offerings or the Enterprise Plus edition.

Architecture

BigQuery data clean rooms are built on a publish and subscribe model of BigQuery data. BigQuery architecture provides a separation between compute and storage, enabling data contributors to share data without having to make multiple copies of the data. The following image is an overview of the BigQuery data clean room architecture:



Data clean room

A *data clean room* is an environment to share sensitive data where raw access is prevented and query restrictions are enforced. Only users or groups that are added as subscribers to a data clean room can subscribe to the shared data. Data clean room owners can create as many data clean rooms as they want in Analytics Hub.

Shared resources

A *shared resource* is the unit of data sharing in a data clean room. The resource must be a BigQuery table or view. As a data contributor, you create or use an existing BigQuery resource in your project that you want to share with your subscribers.

Listings

A *listing* is created when a data contributor adds data into a data clean room. It contains a reference to the data contributor's shared resource along with descriptive information that helps subscribers use the data. As a data contributor, you can create a listing and include information such as a description, sample queries, and links to documentation for your subscribers.

Linked datasets

A *linked dataset* is a read-only BigQuery dataset that serves as a symbolic link to all data in a data clean room. When subscribers query resources in a linked dataset, data from the shared resources is returned, satisfying analysis rules set by the data contributor. As a subscriber, a linked dataset is created inside your project when you subscribe to a data clean room. No copy of the data is created, and subscribers can't see certain metadata, such as view definitions.

Analysis rules

As a data contributor, you configure *analysis rules* on the resources that you share in the data clean room. Analysis rules prevent raw access to underlying data and enforce query restrictions. For example, data clean rooms support the aggregation threshold analysis rule, which lets subscribers analyze data only through aggregation queries.

Data egress controls

Data egress controls are automatically enabled to help prevent subscribers from copying and exporting raw data from a data clean room. Data contributors can configure additional controls to help prevent the copy and export of query results that are obtained by the subscribers.

LAB Section : Hands-on on Analytics Hub (Data Clean Room) capabilities

Steps as Data Publisher :

The Data Publisher in this case is the FraudFix technology. They are providers of data prediction results and model prediction explainability.

1. Create a dataset: ***ml_datasets_clean_room*** which is for the Authorized View. Authorized View is always recommended over table for enforcing the privacy policy. Note how one of the columns are declared as private and put a limit on the lower limit on the aggregated results.
   The dataset should be in the same region as `bootkon_raw_zone dataset that` Dataplex has created before.

   From cloud shell, make sure you set your project ID, then create a BigQuery dataset named ***ml_datasets_clean_room*** in the ***us-central1*** region.

Linux command line : Create a BigQuery dataset: call it ml_datasets_clean_room in US multi region
DATASET_NAME='ml_datasets_clean_room' *bq --location=us-central1 mk -d * *--description "Shared ml dataset " * *$DATASET_NAME*

>>> you can ignore the warning ; warnings.warn("urllib3 ({}) or chardet ({})/charset_normalizer ({}) doesn't match a supported "

2. Define an aggregation threshold analysis rule for a view :
An aggregation threshold analysis rule enforces the minimum number of distinct entities that must be present in a dataset, so that statistics on that dataset are included in the results of a query.
When enforced, the aggregation threshold analysis rule groups data across dimensions, while ensuring the aggregation threshold is met. It counts the number of distinct privacy units (represented by the privacy unit column) for each group, and only outputs the groups where the distinct privacy unit count satisfies the aggregation threshold.
A view that includes this analysis rule can also include the joint restriction analysis rule.

You can define an aggregation threshold analysis rule for a view in a data clean room or with the following statement:

BigQuery SQL : Create shared View
- *Replace your-project-id with your project id*

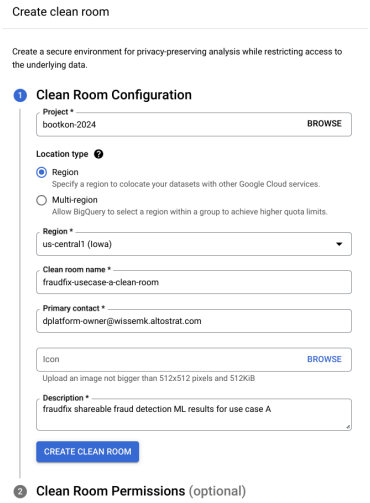
```
CREATE OR REPLACE  VIEW your-project-id.ml_datasets_clean_room.data_prediction_shared
OPTIONS(
 privacy_policy= '{"aggregation_threshold_policy": {"threshold": 1, "privacy_unit_column": "service_account_email"}}'
)
AS ( SELECT * EXCEPT (transaction_id) FROM `your project id.bootkon_raw_zone.data_prediction` );
```

- THRESHOLD : The minimum number of distinct privacy units that need to contribute to each row in the query results. If a potential row doesn't satisfy this threshold, that row is omitted from the query results.
- PRIVACY_UNIT_COLUMN
: Represents the privacy unit column. A privacy unit column is a unique identifier for a privacy unit. A privacy unit is a value from the privacy unit column that represents the entity in a set of data that is being protected.
You can use only one privacy unit column, and the data type for the privacy unit column must be groupable.
The values in the privacy unit column cannot be directly projected through a query, and you can use only analysis rule-supported aggregate functions to aggregate the data in this column.

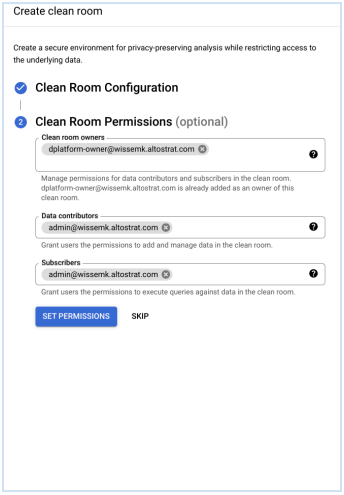
3. Replace <your project id> with your project ID and try the following query in BigQuery without specifying the without an aggregation threshold
SELECT * FROM `<your project id>.ml_datasets_clean_room.data_prediction_shared` LIMIT 1000
Note the error ;
> You must use SELECT WITH AGGREGATION_THRESHOLD for this query because a privacy policy has been set by a data owner.

4. Got to analytics hub and click on ***create clean room***

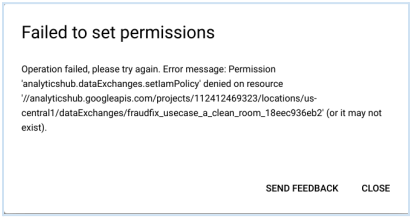
5. Create a Data Clean room called ***fraudfix-usecase-a-clean-room*** in the same region as the ***ml_datasets_clean_room*** dataset (typically us-central1). For the primary contact, use your GCP email address provided to you. For the description, you can use *'Fraudfix shareable fraud detection ML results for use case A'*. Click on create clean room.



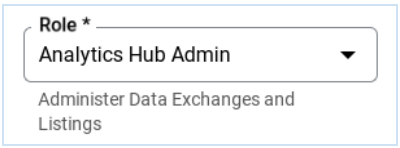
6. Add your GCP email address in the clean room owner field. Add the ***subscriber*** GCP group member email address in both data contributors and subscribers fields, then click on set permissions.



7. Notice the failed permissions



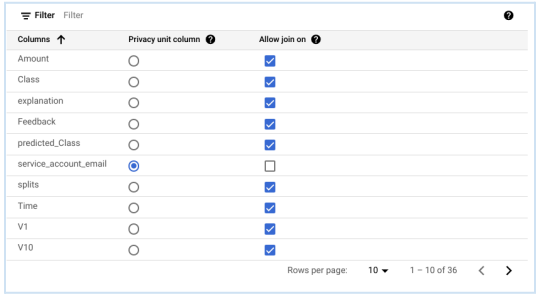
8. After adding the ***Analytics Hub Admin*** role to ***your GCP user***,

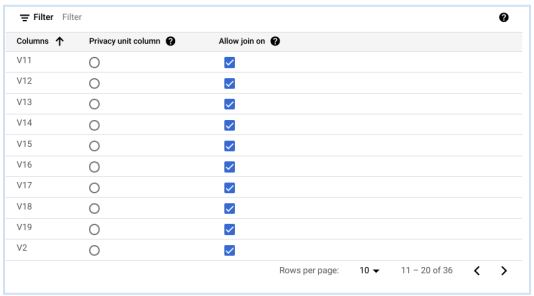


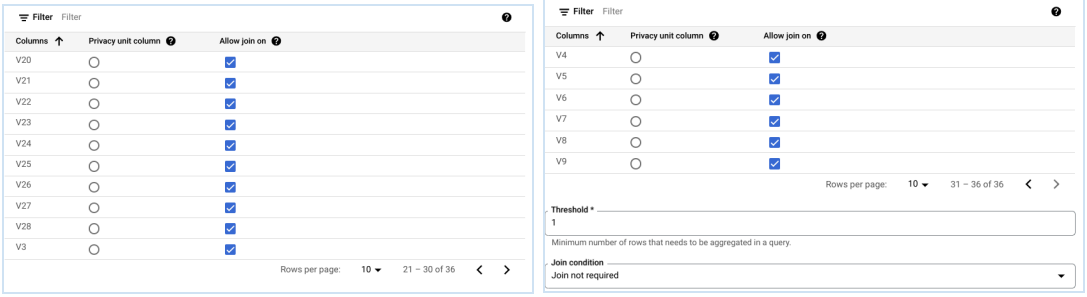
9. Try setting permissions again in step 6. Now, the permissions should be set correctly.

10. In the clean room, add data. Specify the dataset name ***<your project id>.ml_datasets_clean_room*** and add the Auth View ***data_prediction_shared***.

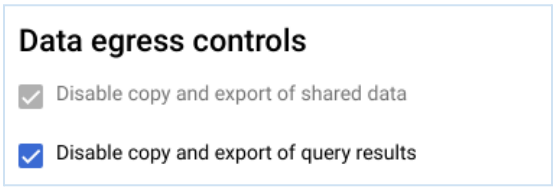
11. Click on next.
12. Notice the privacy unit column is auto detected by the analytics hub.
13. Let ‘s allow the subscribers to join data on all columns except the service account email which is a PII data.



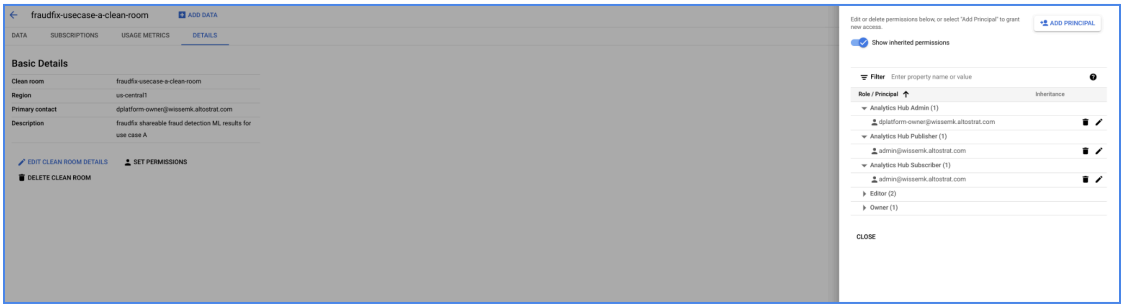




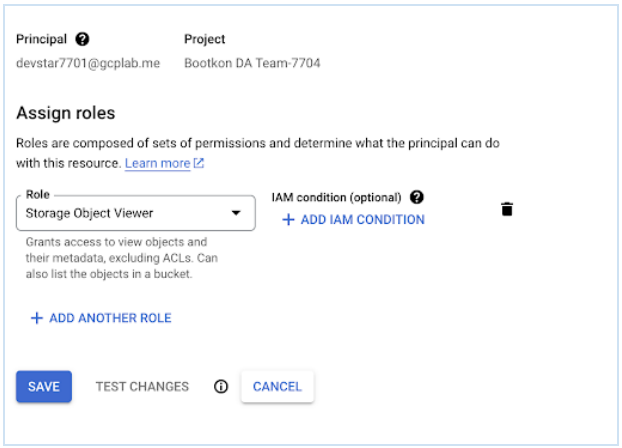
14. Choose the join condition not required.
15. ***Data egress controls*** : Notice you can also disable copy and export of query results. Data egress controls are automatically enabled to help prevent subscribers from copying and exporting raw data from a data clean room. Data contributors can configure additional controls to help prevent the copy and export of query results that are obtained by the subscribers.



16. Review and click on Add data
17. Review the clean room you just created. Especially those who are allowed to subscribe to it. You can always add new principals when needed.



18. Since the table you want to share is in ***BigLake*** table format, grant the '***Storage Object Viewer'*** role to the ***subscriber*** email address. Go to IAM and perform the steps

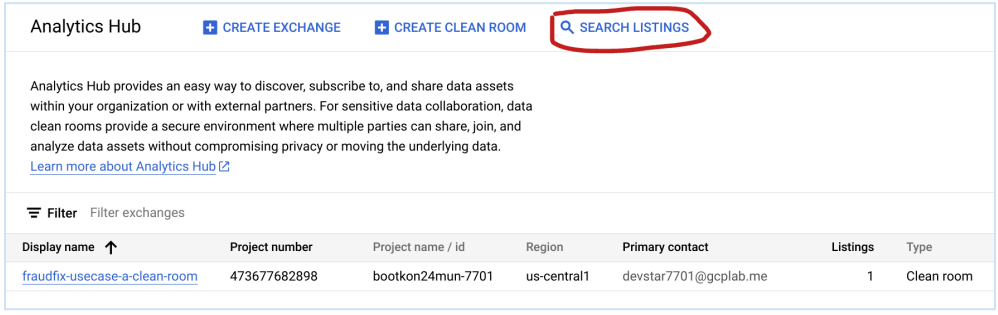


Steps as Data Subscriber :

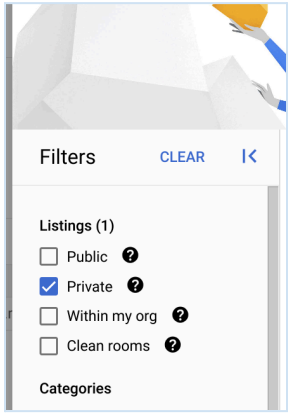
The Data Subscriber in this case is FraudFix's customer. The customer is the owner of the original PCA dataset provided to FraudFix.

1. Go to the ***Analytics Hub*** section in BigQuery and search for Listings by '***Clean rooms***'. The listings shared with you by other group members might take a few minutes to appear.

- Click on ***SEARCH LISTINGS***

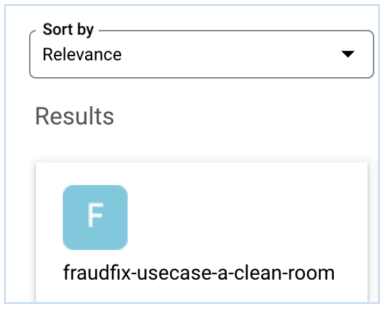


- Then check ***Private Listings*** box from the Filters menu

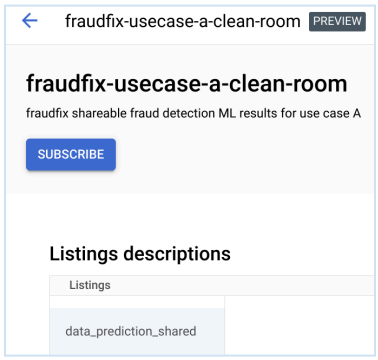


- The results will show the clean rooms shared with you by the other team members.

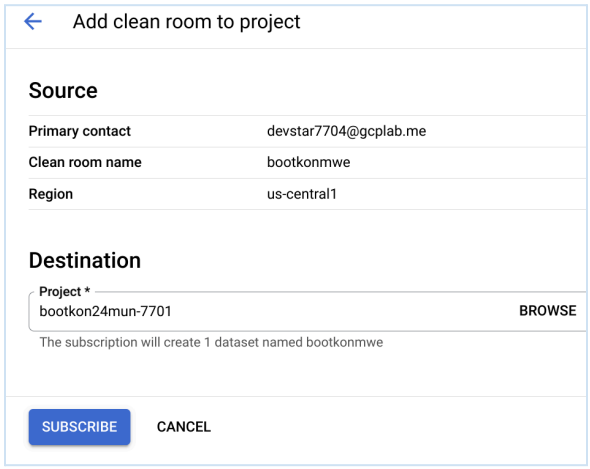
For the remaining steps, we will be working with ***only one*** clean room shared in your listings, so choose ***any*** of the shared clean rooms available to you.



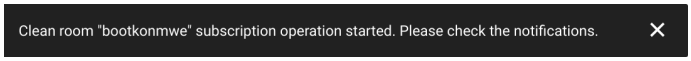
- Now click on the clean room and click on ***SUBSCRIBE***



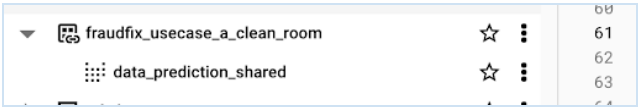
- Add the shared dataset to your BigQuery project id environment by clicking again on ***SUBSCRIBE.*** The destination project should be your project id.



- Notice the message



2. Go to BigQuery Studio, Notice the data clean room that the other team members have just shared with you.



3. Try to select * from the shared table. Note the error;
   You must use SELECT WITH AGGREGATION_THRESHOLD for this query because a privacy policy has been set by a data owner.
4. Try to run a simple aggregation but it will fail to do so as the SQL query must be started with SELECT WITH AGGREGATION_THRESHOLD
5. Run the following query to analyze the results when the predicted class value is different from the actual class value. Replace the ***<your gcp project id>*** *by your actual project id.*

BigQuery SQL : Analyze the results when the predicted class value is different from the actual class value

```
SELECT class, classes, scores from (
SELECT WITH AGGREGATION_THRESHOLD
 class,
 ARRAY(
  SELECT  AS STRUCT classes, scores
  FROM UNNEST(predicted_class.classes) classes WITH OFFSET AS pos
  JOIN UNNEST(predicted_class.scores) scores WITH OFFSET AS pos2
  ON pos = pos2
  ORDER BY scores DESC
  LIMIT 1
 )[OFFSET(0)].*,
FROM
 `<your gcp project id>.fraudfix_usecase_a_clean_room.data_prediction_shared`
GROUP BY class, classes, scores
)
where class <> classes
order by scores desc
```

6. *Try to include the privacy column in your query. For example ;* Replace the ***<your gcp project id>*** *by your actual project id.*

BigQuery SQL : As a data consumer , try to query PII field

```
SELECT service_account_email, class, classes, scores from (
SELECT WITH AGGREGATION_THRESHOLD
 class,
 service_account_email,
 ARRAY(
  SELECT  AS STRUCT classes, scores
```

```
  FROM UNNEST(predicted_class.classes) classes WITH OFFSET AS pos
  JOIN UNNEST(predicted_class.scores) scores WITH OFFSET AS pos2
  ON pos = pos2
  ORDER BY scores DESC
  LIMIT 1
 )[OFFSET(0)].*,
FROM
 `<your gcp project id>.fraudfix_usecase_a_clean_room.data_prediction_shared`
GROUP BY  service_account_email, class, classes, scores
)
where class <> classes
order by scores desc
```

7. Note the error : You cannot GROUP BY privacy unit column when using SELECT WITH AGGREGATION_THRESHOLD
8. Run the following SQL query to know which ***attributes (or features)*** are influencing the model's decision on flagging transactions as fraudulent. Replace <your gcp project id> with your actual project ID.

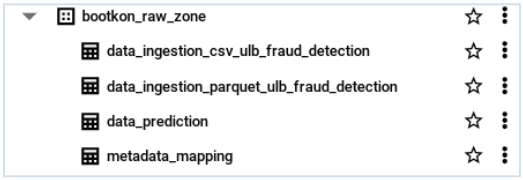
BigQuery SQL : Find the most influential attributes to the model decision

```
WITH RankedPredictions AS (
 SELECT WITH AGGREGATION_THRESHOLD
  class,
  Time,
  ARRAY(
   SELECT AS STRUCT classes, scores
   FROM UNNEST(predicted_Class.classes) classes WITH OFFSET AS pos
   JOIN UNNEST(predicted_Class.scores) scores WITH OFFSET AS pos2
   ON pos = pos2
   ORDER BY scores DESC
   LIMIT 1
  )[OFFSET(0)].*
 FROM
  `<your gcp project id>.fraudfix_usecase_a_clean_room.data_prediction_shared`
  GROUP BY Time, class, classes, scores
),
FilteredRankedPredictions AS (
 SELECT
  Time,
  class,
  classes AS predicted_class,
  scores AS predicted_score
 FROM
  RankedPredictions
```

```
WHERE
  classes = '1'
),
AttributionAverages AS (
SELECT WITH AGGREGATION_THRESHOLD
  AVG(ABS(attribution.featureAttributions.Time)) AS Avg_Time_Attribution,
  AVG(ABS(attribution.featureAttributions.V1)) AS Avg_V1_Attribution,
  AVG(ABS(attribution.featureAttributions.V2)) AS Avg_V2_Attribution,
 AVG(ABS(attribution.featureAttributions.V3)) AS Avg_V3_Attribution,
 AVG(ABS(attribution.featureAttributions.V4)) AS Avg_V4_Attribution,
 AVG(ABS(attribution.featureAttributions.V5)) AS Avg_V5_Attribution,
 AVG(ABS(attribution.featureAttributions.V6)) AS Avg_V6_Attribution,
 AVG(ABS(attribution.featureAttributions.V7)) AS Avg_V7_Attribution,
 AVG(ABS(attribution.featureAttributions.V8)) AS Avg_V8_Attribution,
 AVG(ABS(attribution.featureAttributions.V9)) AS Avg_V9_Attribution,
 AVG(ABS(attribution.featureAttributions.V10)) AS Avg_V10_Attribution,
 AVG(ABS(attribution.featureAttributions.V11)) AS Avg_V11_Attribution,
 AVG(ABS(attribution.featureAttributions.V12)) AS Avg_V12_Attribution,
 AVG(ABS(attribution.featureAttributions.V13)) AS Avg_V13_Attribution,
 AVG(ABS(attribution.featureAttributions.V14)) AS Avg_V14_Attribution,
 AVG(ABS(attribution.featureAttributions.V15)) AS Avg_V15_Attribution,
 AVG(ABS(attribution.featureAttributions.V16)) AS Avg_V16_Attribution,
 AVG(ABS(attribution.featureAttributions.V17)) AS Avg_V17_Attribution,
 AVG(ABS(attribution.featureAttributions.V18)) AS Avg_V18_Attribution,
 AVG(ABS(attribution.featureAttributions.V19)) AS Avg_V19_Attribution,
 AVG(ABS(attribution.featureAttributions.V20)) AS Avg_V20_Attribution,
 AVG(ABS(attribution.featureAttributions.V21)) AS Avg_V21_Attribution,
 AVG(ABS(attribution.featureAttributions.V22)) AS Avg_V22_Attribution,
 AVG(ABS(attribution.featureAttributions.V23)) AS Avg_V23_Attribution,
 AVG(ABS(attribution.featureAttributions.V24)) AS Avg_V24_Attribution,
 AVG(ABS(attribution.featureAttributions.V25)) AS Avg_V25_Attribution,
 AVG(ABS(attribution.featureAttributions.V26)) AS Avg_V26_Attribution,
 AVG(ABS(attribution.featureAttributions.V27)) AS Avg_V27_Attribution,
  AVG(ABS(attribution.featureAttributions.V28)) AS Avg_V28_Attribution,
  AVG(ABS(attribution.featureAttributions.Amount)) AS Avg_Amount_Attribution
FROM
  `<your gcp project id>.fraudfix_usecase_a_clean_room.data_prediction_shared` DP
JOIN
  FilteredRankedPredictions FRP
ON
  DP.Time = FRP.Time
CROSS JOIN
  UNNEST(DP.explanation.attributions) as attribution
WHERE
  FRP.class = '1'
)
```

```
SELECT * FROM AttributionAverages
```

9. Let's map the results of the previous query with our secret metadata PCA mapping table to understand which attributes are heavily influencing the model's fraudulent decisions. Notice we already have a Biglake table created by Dataplex under the ***bootkon_raw_zone*** dataset called ***metadata_mapping***.
Using the previous SQL statement, you find out the most influential attributes ; for example V14.
This table should be accessible only by the customers of FraudFix and not by FraudFix employees because it can be used to reverse PCA and access customer private information.



10. Query the metadata table ***"metadata_mapping"*** and take note of the meanings and descriptions of the most influential V* attributes (both higher value and lower value attributes). For example, ***V14*** is the most influential attribute for ML decisions. ***V14*** corresponds to the dimensional PCA space attribute for ***"Dispute and Chargeback Frequency".*** It measures the frequency of disputes and chargebacks, which can be a direct indicator of customer dissatisfaction or fraudulent transactions. Remember that when FraudFix received the dataset from their customers, they did not know the meanings of the V* columns and their values. FraudFix does not have access to the PCA metadata table. However, as a subscriber (FraudFix customer), you have access to the PCA metadata.

🥳🥳Congratulations on completing Lab 6!
You can now move on to Lab 7 for further practice.🥳🥳