FraudFix Use Case

Data Governance with Dataplex

Duration: 60 Minutes

Authors: Wissem Khlifi
Date : April 1, 2024

Architecture Diagram:
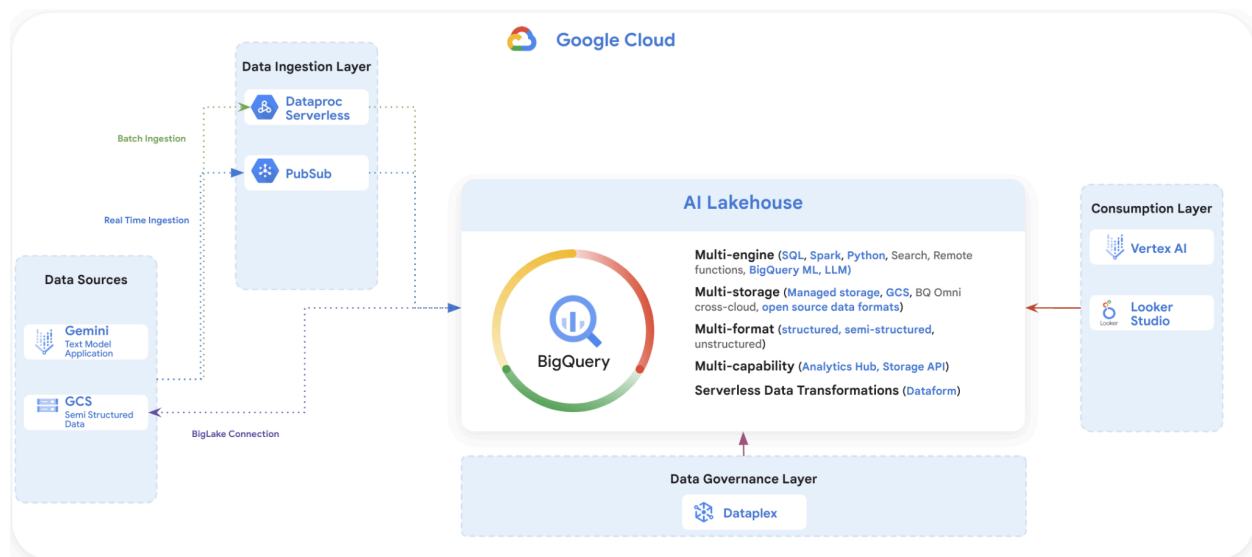
Goal of the Lab:

- Understand Dataplex product capabilities
- Leverage Dataplex features to better understand, govern your data and metadata.
- Build data quality checks on top of the fraud detection prediction results.

Lab Dependencies:

We have made this Lab 5 independent from the previous machine learning Lab 4. During previous machine learning with Vertex AI lab, we discussed that batch prediction jobs may take more than *2 hours* to complete. Therefore, we made available the results of the job prediction in parquet files here. Ensure these files are copied into your GCS bucket *<gcp project id>-bucket*. Please note however that this Lab 5 is dependent on *Lab1* and *Lab 2*.

READING Section : What Dataplex is and what problems it solves

What problems does Dataplex solve?

Organization

With Dataplex you can organize different data assets, from different projects under new organizational concepts of Lakes and Zones. Organization is logical only and does not require any data movement. Dataplex supports managing datasets in BigQuery and GCS buckets. You can use lakes to define your organizational boundary or regional boundary (e.g. marketing lake/sales lake Or US lake/ UK lake etc), while zones can be used to group the data logically or by use cases (e.g. raw_zone/curated_zone or analytics_zone/data_science_zone).
Dataplex can also be used to build a data mesh architecture with decentralized data ownership among domain data owners.

Security - GCS / BQ

With Dataplex you can apply data access permissions using IAM groups across multiple buckets and BQ datasets by granting permissions at a lake or zone-level. Dataplex will do the heavy lifting of propagating desired policies and updating access policies of the buckets/datasets that are part of that lake or data zone. Dataplex will also apply those permissions to any new buckets/datasets that get created under that data zone. This takes away the need to manually manage individual bucket permissions and also provides a way to automatically apply permissions to any new data added to your lakes.

Note that the permissions are applied in "Additive" fashion. I.e. Dataplex does not replace the existing permissions when pushing down permissions. Dataplex also provides "exclusive" permission push down as an opt-in feature.
Discovery [semi structured and structured data].
You can configure discovery jobs in Dataplex that can sample data on GCS, infer its schema, and automatically register it with Data Catalog so you can easily search and discover the data you have in your lakes.

In addition to registering metadata with Data Catalog, for data in CSV, JSON, AVRO, ORC, and Parquet formats, the discovery jobs also register technical metadata, including hive-style partitions, with a managed Hive metastore (Dataproc Metastore) & as external tables in BigQuery(BQ). Discovery jobs can be configured to run on a schedule to discover any new tables or partitions. For new partitions, discovery jobs incrementally scan new data, check for data and schema compatibility, and register only compatible schema to the Hive metastore/ BQ so that your table definitions never go out of sync with your data.

Actions - Profiling, Quality, Lineage, Discovery

Dataplex has the capability to profile data assets (BigQuery tables) , auto detect data lineage for BigQuery transformations. You can also use Dataplex for data discovery across GCS, BigQuery, Spanner, PubSub, Dataproc metastore, Bigtable and Vertex AI models. Dataplex automatic data quality, which lets you define and measure the quality of your data. You can automate the scanning of data, validate data against defined rules, and log alerts if your data doesn't meet quality requirements. You can manage data quality rules and deployments as code, improving the integrity of data production pipelines.

LAB Section : Hands-on on Dataplex capabilities

[LAB] Create Dataplex Lake

1. Enable the Dataplex, Dataproc, Dataproc Metastore, Data Catalog, BigQuery, and Cloud Storage. APIs. **(you can skip this step if you completed LAB 1)**
2. Make sure you have the predefined roles *roles/dataplex.admin* or *roles/dataplex.editor* granted to you so that you can create and manage your lake. **(you can skip this step if you completed LAB 1)**
3. Go to Dataplex in the Google Cloud console.
4. Navigate to the Manage view.
5. Click Create Lake.

6. Enter a Display name. For example: bootkon-lake
7. The lake ID is automatically generated for you. If you prefer, you can provide your own ID.
8. Optional: Enter a Description. For Example: Dataplex Lake for bootkon data assets
9. Specify the Region where the GCS buckets and BigQuery datasets were created during previous Labs. If you have followed the previous Labs, it should be us-central1. Ensure that the region is consistent with the locations used in prior steps.

10. Optional: Add labels to your lake. For example, use location for the key and berlin for the value.
11. Lets skip the metastore creation for now and click on create.
12. The creation should take 2-3 minutes to finish.

Display name ↑	Assets	Status	Resources requiring action	Region	Last modified	Labels	
BOOTKON-LAKE	0	✅ Active	0	us-central1 (Iowa)	April 17, 2024	location : berlin	⋮

[LAB] ADD Dataplex Zones

Data zones are named entities within a Dataplex lake. They are logical groupings of unstructured, semi-structured, and structured data, consisting of multiple assets, such as Cloud Storage buckets, BigQuery datasets, and BigQuery tables.

A lake can include one or more zones. While a zone can only be part of one lake, it may contain assets that point to resources that are part of projects outside of its parent project.

You can select configurations for a zone in Dataplex. There are two types of zones that you can choose from: raw and curated zones. (For explanation of raw and curated zones refer to the "Explanation of Raw and Curated Zones" section of the appendices).

We will add 2 zones; one for raw zone and another one for curated zone.

1. Click on the bootkon-lake lake you just created.
2. In the **Zones** tab, click + **Add zone**.
3. Enter a **Display name** for your zone. For example ; bootkon-raw-zone
4. Click the **Type** drop-down. Choose **Raw Zone**. Learn more about underlined supported zone types.
5. Optional: Enter a description. For example, Dataplex zone for bootkon raw data assets
6. Under **Data locations**, select either **Regional** or **Multi-regional**. The region has to match where the GCS buckets and BigQuery datasets were created during previous Labs. If you have followed the previous Labs, it should be us-central1.
7. Add labels to your zone. For example, use **zone** for the key and **raw** for the value.
8. Enable metadata discovery, which allows Dataplex to automatically scan and extract metadata from the data in your zone. Let's leave the default settings.
   a. Expand the **Discovery settings** submenu.
   b. Make sure **Enable metadata discovery** is selected.
   c. Optional: Under **Include patterns**, list the files to include in the discovery scans.
   d. *Important*: Under Exclude patterns, list the files to exclude from the discovery scans. If you enter both include and exclude patterns, exclude patterns are applied first. Exclude the source code files by specifying ***/src/***.

<blockquote>

e. Click the **Repeats** drop-down and select a frequency.

f. Click the **Timezone** drop-down and select a timezone.

g. If under **Repeats** you selected **Custom**, under **Schedule**, enter a job schedule. Otherwise, the **Schedule** value is automatically filled for you.

</blockquote>

9. Click **Create**.
10. When the zone creation succeeds, the zone automatically enters an active state. If it fails, then the lake is rolled back to its previous state.
11. After you create your zone, you can map data stored in Cloud Storage buckets and BigQuery datasets as assets in your zone.
12. Repeat the same steps from 1 to 11 but this time, change the display name to bootkon-curated-zone and choose Choose **Curated Zone** for the Type. You might also change the label and description values.
13. The creation should take 2-3 minutes to finish.

[LAB] ADD Zone Data Assets

Lets map data stored in Cloud Storage buckets and BigQuery datasets as assets in your zone.

1. Click on bootkon-raw-zone
2. Click on + ADD ASSETS
3. Click on ADD AN ASSET
4. Choose storage bucket
5. Display name : bootkon-gcs-raw-asset
6. Optionally add a description
7. Browse the bucket name and choose the bucket created in LAB 1. If you followed the instructions, it should be named ***<your project id>-bucket***.
8. Select the bucket
9. Let's skip upgrading to the managed option. When you upgrade a Cloud Storage bucket asset, Dataplex removes the attached external tables and creates BigLake tables. We have already created in LAB 2 biglake table so this option is not necessary.
10. Optionally add a label
11. Click on continue
12. Leave the discovery setting to be inherited by the lake settings we have just created during lake creation steps. Click on continue.
13. Click on submit.

Lets add another data assets but for the bootkon-curated-zone

14. Click on bootkon-curated-zone

15. Click on + ADD ASSETS

16. Click on ADD AN ASSET

17. Choose BigQuery Dataset

18. Display name : bootkon-bq-curated-asset

19. Optionally add a description

20. Browse the BigQuery Dataset and choose the dataset created in LAB 1. If you followed the instructions, it should be named *ml_datasets*.

21. Select the BigQuery Dataset

22. Optionally add a label

23. Click on continue

24. Leave the discovery setting to be inherited by the lake settings we have just created during lake creation steps. Click on continue.

25. Click on submit.

[LAB] Explore the data assets with Dataplex Search

During this lab go to the Search section of the Dataplex and search for the lakes, zones and assets you just created. Spend 5 minutes before moving to the next LAB.

[LAB] Explore Biglake object tables created automatically by Dataplex in BigQuery

As a result of the data discovery , notice a new BigQuery dataset created called "bootkon_raw_zone". New Biglake tables were automatically created by Dataplex discovery jobs. During the next sections of the labs, we will be using the data_prediction biglake table.

During previous machine learning with Vertex AI lab, we discussed that batch prediction jobs may take more than **2 hours** to complete. Therefore, we made available the results of the job prediction in parquet files [here](). Ensure these files are copied into your GCS bucket **<gcp project id>-bucket**.

[LAB] Exploring Data Lineage

During previous labs, we have created a series of data transformations in BigQuery using Dataform.
In this lab, we can discover the type of transformations and the entities involved as well as who initiated them through data lineage graphs.
Go to the Search section of the Dataplex.
We can tailor the search of the sentiment inference BigQuery table by add more filters to the search,
Under Filters section and under Systems, Choose BigQuery

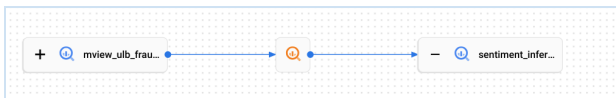Choose Table under Data Types

Choose ml_datasets under Datasets section
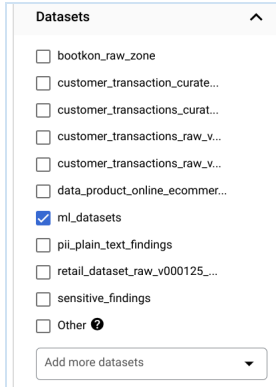
Click on sentiment inference table
Under Lineage section, explore the lineage graph

[LAB Data Profiling]

Dataplex data profiling lets you identify common statistical characteristics of the columns in your BigQuery tables. This information helps you to understand and analyze your data more effectively.
Information like typical data values, data distribution, and null counts can accelerate analysis. When combined with data classification, data profiling can detect data classes or sensitive information that, in turn, can enable access control policies.
Dataplex also uses this information to recommend rules for data quality checks.

Dataplex lets you better understand the profile of your data by creating a data profiling scan.

The following diagram shows how Dataplex scans data to report on statistical characteristics.
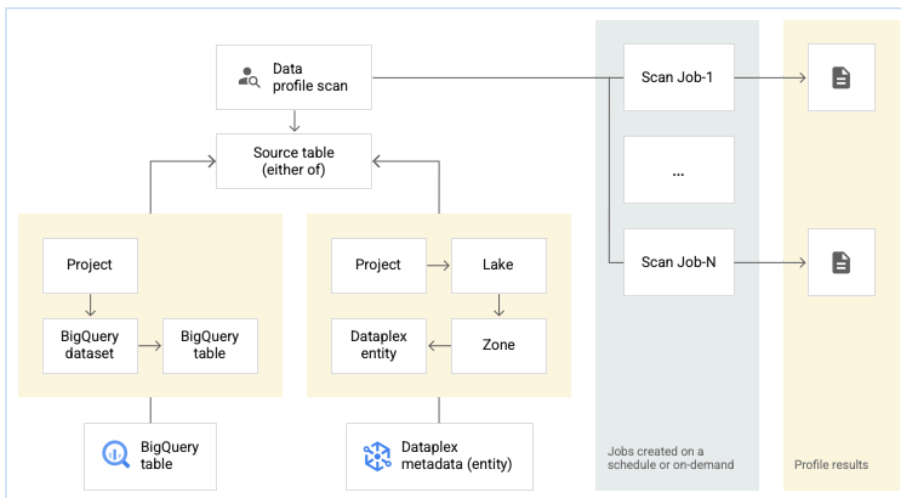
Configuration options

This section describes the configuration options available for running data profiling scans.

Scheduling options

You can schedule a data profiling scan with a defined frequency or on demand through the API or the Google Cloud console.

Scope

As part of the specification of a data profiling scan, you can specify the scope of a job as one of the following options:

- Full table: The entire table is scanned in the data profiling scan. Sampling, row filters, and column filters are applied on the entire table before calculating the profiling statistics.
- Incremental: Incremental data that you specify is scanned in the data profile scan. Specify a Date or Timestamp column in the table to be used as an increment. Typically, this is the column on which the table is partitioned. Sampling, row filters, and column filters are applied on the incremental data before calculating the profiling statistics.

Filter data

You can filter data to be scanned for profiling by using row filters and column filters. Using filters helps you reduce the execution time and cost, and exclude sensitive and unuseful data.

- Row filters: Row filters let you focus on data within a specific time period or from a specific segment, such as region. For example, you can filter out data with a timestamp before a certain date.
- Column filters: Column filters lets you include and exclude specific columns from your table to run the data profiling scan.

Sample data

Dataplex lets you specify a percentage of records from your data to sample for running a data profiling scan. Creating data profiling scans on a smaller sample of data can reduce the execution time and cost of querying the entire dataset.

Lab Instructions

1. Go to Profile section in Dataplex
2. Click on +CREATE DATA PROFILE SCAN
3. Display Name: bootkon-dprofile-fraud-prediction for example
4. Optionally add a description. For example, data profile scans for fraud detection predictions
5. Leave the "browse within dataplex lakes" option turned off
6. Click on browse to filter on data_prediction bigquery table.

Select table

To create data scans on bigquery external table, please make sure your service account has storage.objects.list permission for the corresponding GCS Bucket.

DISMISS

	Name	Description	Project
○	gcp_billing_export_resource_v1_011FC3_F6B233_EAC271 Dataset: all_billing_data	-	data-product-analytic-ecommrce
○	ulb_fraud_detection_prediction Dataset: ml_datasets	-	bootkon-2024
○	data_ingestion_parquet_ulb_fraud_detection	-	bootkon-2024
○	data_prediction	-	bootkon-2024
◉	data_prediction Dataset: bootkon_raw_zone	-	bootkon-2024
○	data_ingestion_csv_ulb_fraud_detection	-	bootkon-2024
○	data_ingestion_parquet_ulb_fraud_detection Dataset: bootkon_raw_zone	-	bootkon-2024
○	data_ingestion_csv_ulb_fraud_detection Dataset: bootkon_raw_zone	-	bootkon-2024
○	ulb_fraud_detection_prepped	-	bootkon-2024
○	sentiment_inference	-	bootkon-2024
○	ulb_fraud_detection	-	bootkon-2024
○	ulb_fraud_detection_parquet	-	bootkon-2024
○	sentiment_inference Dataset: ml_datasets	-	bootkon-2024
○	ulb_fraud_detection_blake Dataset: ml_datasets	-	bootkon-2024
○	mview_ulb_fraud_detection	-	bootkon-2024

Rows per page: 15 ▾   1 – 15 of many   ‹ ›

7. Select data_prediction bigquery table
8. Choose "entire data" as scope of the data profiling job
9. Choose All data on sampling size
10. Turn on publishing option
11. Choose on demand schedule
12. Click on continue, leave the rest as default and click on create.
13. It would take a couple of minutes for the profiling to show up on the console.


Data Profiling   ⊞ CREATE DATA PROFILE SCAN   ⊞ CREATE MULTIPLE PROFILE SCANS   ⟳ REFRESH

Analyze the profile of your Dataplex managed data by configuring and scheduling checks over your data.

⇄ Filter

	Display name ↑	Last run	Labels	Table name	Incremental column
☐	BOOTKON-DPROFILE-FRAUD-PREDICTION		None	data_prediction	

14. Click on the bootkon-dprofile-fraud-prediction profile and click on RUN NOW.
15. Click on Job ID and monitor the job execution.
16. Notice what the job is doing.
17. The Job should succeed in less than 10 minutes.
18. Explore the data profiling results of the CLASS column name. We have less than 0.1% of fraudulent transactions. Also notice that predicted_class of type RECORD were not fully profiled, only the percentage of null and unique values were correctly profiled. Refer to the supported data types here.
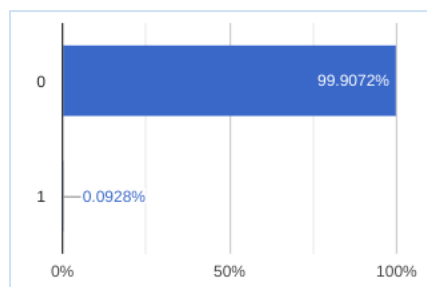
19. As they train further and continuously the fraud detection ML models, data professionals would like to set up an automatic check on data quality and be notified when there are huge discrepancies between predicted_class and CLASS values. This is where Dataplex data quality could help the team.

[LAB] Setup Data Quality Jobs

After setting up the data profiling scan we have seen that we still have no clear visibility on fluctuation between predicted classes vs actual CLASS ratio. Our goal is to have a percentage of matched values between CLASS and predicted classes more than 99.99 %. Any lower percentage would indicate that we would have to further train the ML model or add more features or use another model architecture.

You can use the following SQL query in BigQuery to check the percentage of matched values between CLASS and predicted classes.

**BigQuery SQL : Check the percentage of matched values between CLASS and predicted classes**
- *Replace your-project-id with your project id*

```
WITH RankedPredictions AS (
SELECT
 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-project-id.bootkon_raw_zone.data_prediction`
)

SELECT
 SUM(CASE WHEN class = CAST(highest_score_class AS STRING) THEN 1 ELSE 0 END) * 100.0 / COUNT(*)  AS
PercentageMatch
FROM (
SELECT
 class,
 classes AS highest_score_class
FROM
 RankedPredictions

)
```

We will set up the Dataplex automatic data quality, which lets you define and measure the quality of your data. You can automate the scanning of data, validate data against defined rules, and log alerts if your data doesn't meet quality requirements. You can manage data quality rules and deployments as code, improving the integrity of data production pipelines.

During the previous lab, We got started by using Dataplex data profiling rule recommendations to drive initial conclusions on areas of attention. Dataplex provides monitoring, troubleshooting, and Cloud Logging alerting that's integrated with Dataplex auto data quality.

Conceptual model
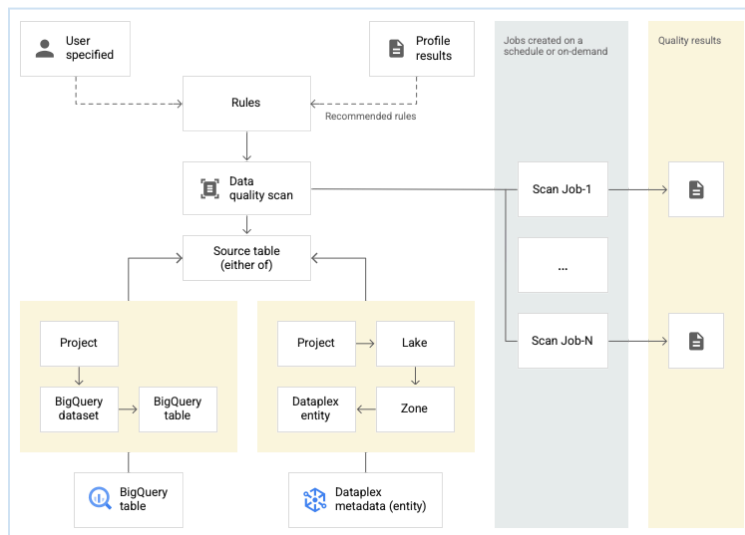
A data scan is a Dataplex job that samples data from BigQuery and Cloud Storage and infers various types of metadata. To measure the quality of a table using auto data quality, you create a DataScan object of type data quality. The scan runs on only one BigQuery table. The scan uses resources in a Google tenant project, so you don't need to set up your own infrastructure.
Creating and using a data quality scan consists of the following steps:

1. Rule definition
2. Rule execution
3. Monitoring and alerting
4. Troubleshooting

Lab Instructions

1. Go to Data Quality section in Dataplex
2. Click on +CREATE DATA QUALITY SCAN
3. Display Name: bootkon-dquality-fraud-prediction for example
4. Optionally add a description. For example, data quality scans for fraud detection predictions
5. Leave the "browse with dataplex lakes" option turned off
6. Click on browse to filter on the *data_prediction* BigQuery table.
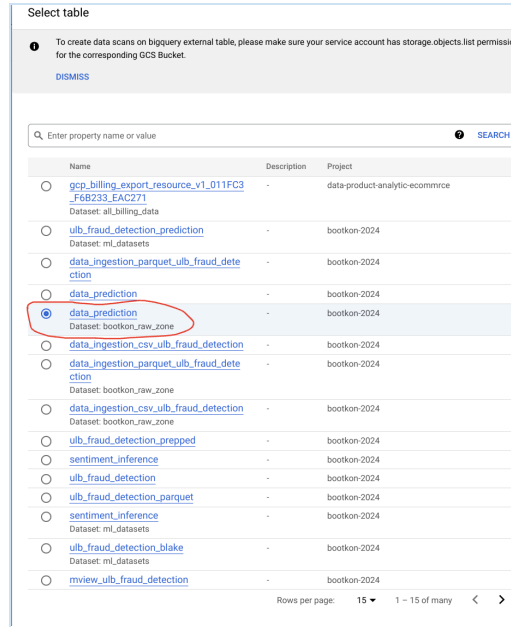
7. Select data_prediction bigquery table
8. Choose "entire data" as scope of the data profiling job
9. Choose All data on sampling size
10. Turn on publishing option
11. Choose on demand schedule
12. Click on continue,
13. Now lets define quality rules, click on ADD RULES > SQL Assertion Rule
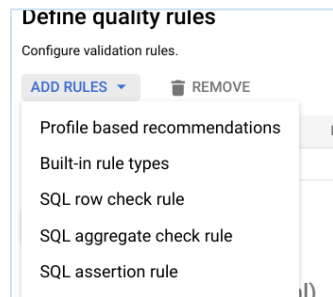
14. Choose Accuracy as dimension
15. Rule name: bootkon-dquality-ml-fraud-prediction
16. Description : regularly check the ML fraud detection prediction quality results
17. Leave the column name empty
18. Provide the following SQL stat*ement. Dataplex will utilize this SQL statement to create a SQL clause of the form **SELECT COUNT(*) FROM** (sql statement) to return success/failure. The assertion rule is passed if the returned assertion row count is **0**.*

BigQuery SQL : Assertion SQL

```
WITH RankedPredictions AS (
 SELECT
  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
  `bootkon-2024.bootkon_raw_zone.data_prediction`
)

SELECT
 SUM(CASE WHEN class = CAST(highest_score_class AS STRING) THEN 1 ELSE 0 END) * 100.0 / COUNT(*)  AS PercentageMatch
FROM (
 SELECT
  class,
  classes AS highest_score_class
 FROM
  RankedPredictions

)
 HAVING PercentageMatch <= 99.99
```

19. Click on ADD
20. Click on Continue

☐	Column name ↑	Rule name	Rule type	Evaluation	Dimension	Parameters	Threshold	Edit
☐	-	bootkon-dquality-ml-fraud-prediction	Sql Assertion Check	Aggregate	Accuracy	WITH RankedPredictions AS (SELECT 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 `bootkon-2024.bootkon_raw_zone.data_prediction`) SELECT SUM(CASE WHEN class = CAST(highest_score_class AS STRING) THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS PercentageMatch FROM (SELECT class, classes AS highest_score_class FROM RankedPredictions) HAVING PercentageMatch >= 99.99	N/A ✏	

21. Run SCAN
22. Monitor the job execution. Notice the job succeeded but the rule failed because our model accuracy percentage on the whole data predicted does not exceed the 99.99% threshold that we set.

Job ID: 06619967-7faa-4d33-b48d-2a9ddca39c60 results ✕

Job details

Job ID	06619967-7faa-4d33-b48d-2a9ddca39c60	Row filter	
Records scanned	276997	Sampling size	All data
Start time	April 17, 2024 at 2:42:30 PM UTC+2	Results exported to	N/A
End time	April 17, 2024 at 2:43:00 PM UTC+2	Results export status	
Scope	Entire data		
Job status	✅ Succeeded		

Dimensions failed

Accuracy	❗ 1 Errors

Rules

Filter | Filter Items   ❓ ▥

Column name ↑	Rule name	Rule type	Status	Evaluation	Dimension	Parameters	Failed rows	Threshold	Query to get failed records	Copy
-	bootkon-dquality-ml-fraud-prediction	Sql Assertion Check	❗ Failed	Aggregate	Accuracy	WITH RankedPredictions AS (SELECT class,	0%	N/A	-- This query executes the SQL expression...	📋 ⌄

<h1 style="text-align:center">🥳🥳Congratulation you have successfully completed LAB 5 🥳🥳</h1>

[OPTIONAL HOMEWORK CHALLENGE LAB] Set alerts in Cloud Logging

Set Alerts in Cloud Logging. You can follow the instructions here;
[OPTIONAL TASK][Home work Challenge LAB] Secure your Lake
Secure the dataplex lake. For instructions, follow the instructions here.
[OPTIONAL TASK][Home work Challenge LAB] Dataplex Glossaries
Create business glossaries. For instructions, follow the instructions here.
[OPTIONAL TASK][Home work Challenge LAB] Dataplex Integrations
Explore assets in the looker studio and scan sensitive data with DLP. However due to environment restrictions, you might not have the privileges to initiate DLP scans. If so, then check with your GCP organization administrator.

Appendices

Explanation of Raw and Curated Zones

Raw Zone
The Raw Zone of a data lake stores unprocessed data as it comes from the source. It's mostly used by automated pipelines and data engineers for data cleansing and transformation. It is recommended to organize the data by source for lifecycle management and billing purposes. Cloud Storage is generally used for batch files, while BigQuery tables are typically used for raw streaming data.

Curated Zone
The Curated Zone is a highly structured and validated layer within a lakehouse, designed for traditional analytics and decision-making by business users. It contains aggregated data and may be cross-joined with other lakehouse data. Accessible via BigQuery datasets or Cloud Storage buckets, the data is easily discoverable and of high quality. While data engineers can update it through pipelines, data scientists and analysts can also modify it using their preferred tools.

Find more information about Lakehouse on Google Cloud in the Whitepaper ; Building the analytics lakehouse on Google Cloud