Forecasting Numbers of Learners with ML Model in GCP

SCS 3760-002 Term Project

Hwarang Kim

Business Case

Forecast Numbers of Learners in 2021Q2

How are numbers of learners trending?

Background

- Members of Professional Association (PA) need to meet Annual Learning Requirements (ALR).
- Members can take ALR courses via PA and other learning institutions.
- Due to the current pandemic, PD team is required to re-forecast numbers of learners in 2021 Q2. The re-forecasted numbers will be provided to Finance and PD team to adjust the current financial plan.

Business Problem

- How are numbers of learners trending?
- Need to forecast numbers of learners in 2021 Q2

Machine Learning (ML) Model

Supervised Binary Classification Model

- Label:
  - 1 = Take ALR courses
  - 0 = Not take ALR courses
- Features
  - Member profiles (age, gender, tenure, employer size, industry, etc.)
  - Member ALR transactions (last 3-years)
  - Please see *Appendix A* for details

Cost-Benefit Analysis: Cloud vs On-premise

Benefits of Cloud computing

- Intensive and scalable computing power for ML/AI
- Only pay for what you use (Opex)
- Agile and low setup cost and effort
- Managed services; no maintenance
- Shared security responsibility model

Costs of On-premise computing

- On-prem computing for ML/AI is often less scalable and high cost
- High capital expenditure (Capex)
- Monitoring, patching and maintaining on-premise infrastructures
- Accountable for entire security responsibility

GCP vs AWS

GCP

- (-) Launched in 2011 and the third place in cloud market (HSBC, PayPal, Home Depot) (please see *Appendix B*)
- (-) 60+ Services
- (+) Lower cost and data access
- (+) Faster and easier deployment
- (+) Suited for the secondary cloud platform
- (+) High-end computing offerings (big query, analytics and machine learning)

AWS

- (+) The first to cloud market since 2006 and the market share leader (Netflix, AirBnB, Samsung)
- (+) 200+ Services
- (-) Higher cost and data access; complex cost structure
- (-) Slower deployment speed

Design of Solution Architecture

The Cloud Solution by Google Cloud Platform (GCP)

Architectural Requirements

- Data integration with external data sources (Salesforce and MS Dynamics365)
- Provision of staging / landing area for external data sources

- Data warehouse capability hosing dim and fact tables/views
- Data storage for external datasets
- Data transformation with SQL query
- Data storage for tables and views for transformed datasets

- Prep dataset for ML
- Train ML models
- Make predictions
- Jupyter notebooks

- Visualize predictions
- Dashboards
- Reports

GCP Resources Review

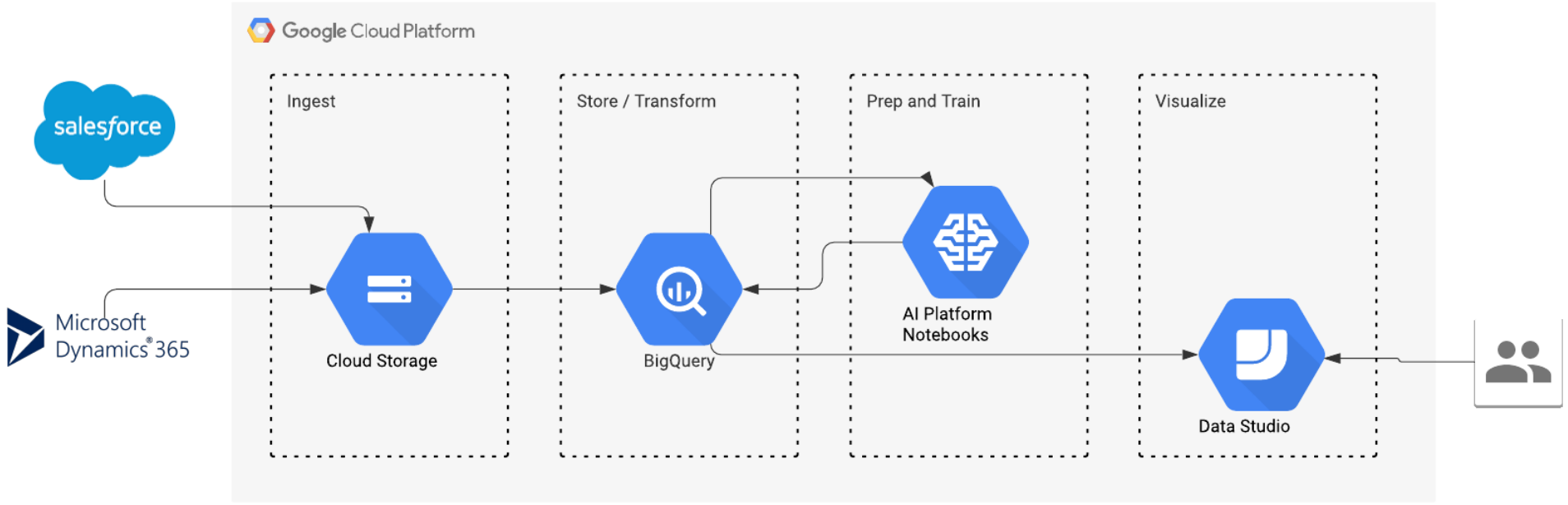
GCP Resources	Ingest / Store / Transform
Cloud SQL: Fully managed relational database.	(-) Required to map schema of external data sources and targe tables. (-) Not suited for staging / landing area for massive external data sources. (+) Relational database; SQL to transform datasets. (+) Moderate cost.
Cloud Spanner: Scalable and Global database with high consistency.	(-) Multi-regional storages are not required for this project. (+) High cost.
Cloud Bigtable: Wide-column NoSQL database used for high-volume database.	(-) Wide-column NoSQL database is not necessary for this project. (-) Suited for IoT, time-series and similar applications. Not required fort this project.
Cloud Firestore: Managed document database which are used when the structure of data vary from one record to another.	(-) Document database is not required for this project.
Cloud Storage:** Object storage system designed for any data type (structured, semi-structured, and unstructured).	(+) Cloud Storage can store any data type (e.g., csv, parquet, etc.). (+) Suited for staging and landing area for transformation and storage (+) Low cost.
BigQuery: Fully managed, petabyte-scale, low-cost analytics data warehouse databases.	(+) Agile to load and export datasets; connected to cloud storage as external tables. (+) Use SQL-like commands to query massive datasets very quickly and create data pipelines. (+) Data warehouse hosting tables and views for analytics and reporting. (+) Low cost.

**Note: Data integration with Salesforce and MS Dynamics need to be further investigated. Please see *Appendix C*.

GCP Resources Review

GCP Resources	Prep / Train / Visualize
AI Platform Notebooks: Managed service that offers an integrated and secure JupyterLab environment for data scientists and ML developers.	(+) Jupyter notebook available with deploying a separate instance for VM. (+) Easy deployment of instances running JupyterLab that come pre-installed with the latest data science and ML frameworks. (+) Load and extract datasets between BigQuery and Cloud Storage
Big Query ML: It enables users of the analytical database to build ML models using SQL and data in BigQuery datasets.	(+) Making ML method available through SQL functions. (+) Not having to import and export datasets for ML.
Cloud AutoML: Machine learning service designed for developers who want to incorporate ML into their applications without having to write Python codes.	(+) Develop and deploy ML model without writing Python codes. (+) Enables developers with limited ML expertise to train models specific to their business needs. (-) High cost (-) Longer training time
Data Studio: Free visualization tool offered by Google.	(+) Google native visualization solution. (-) Low loading speed if use blended data/ high quantity of data/ a lot of calculated metrics. (-) Does not have all required connectors. Will be required to use some third-party tools.

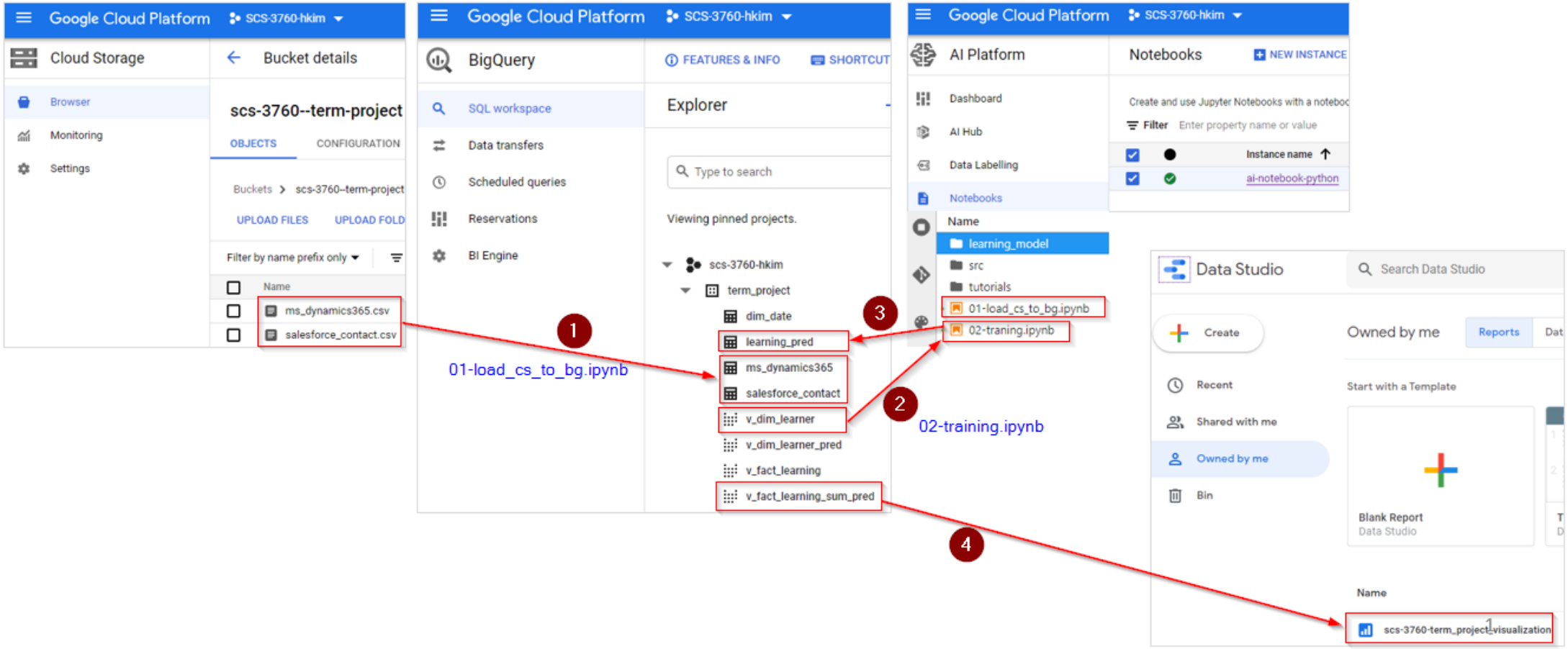
Cloud Architecture



Implementation of Solution

Deploying GCP Resources (Cloud Storage, Big Query, AI Platforms)

Solution Implementation



AI Platform: Jupyter Notebook

Load Salesforce Contact from Cloud Storage to BigQuery

- Download csv file from Cloud Storage
  - Google storage library
- Load table from dataframe
  - Google bigquery library

```
# ========================
# 1. Salesforce Contact
# ========================
# download from cloudstorage
from google.cloud import storage
from io import StringIO
import pandas as pd

client = storage.Client()
bucket = client.get_bucket('scs-3760--term-project')

# download csv from cloudstorage
blob = bucket.get_blob(f"salesforce_contact.csv")
bt = blob.download_as_string()
s = str(bt, "utf-8")
s = StringIO(s)
df_cs = pd.read_csv(s)

# load to bigquery
from google.cloud import bigquery
client = bigquery.Client(location="US")

# drop table salesforce_contact
sql = "drop table if exists `scs-3760-hkim.term_project.salesforce_contact`"
client.query(sql, location="US")

bg_dataset = client.dataset('term_project')
table_ref = bg_dataset.table("salesforce_contact")
job = client.load_table_from_dataframe(df_cs, table_ref, location="US")

job.result()  # Waits for table load to complete.
print("Loaded dataframe to {}".format(table_ref.path))
```

BigQuery: Transformation

Creating a view by transforming transactional data to columns (features for ML model)

View name: v_fact_learning

Data sources

- ms_dynamics365: transactional table including learning activities
- dim_date: dimensional table for date aggregate

View info

View ID	scs-3760-hkim:term_project.v_fact_learning
Created	18 Apr 2021, 16:32:11
Last modified	18 Apr 2021, 17:40:10
View expiry	Never
Use Legacy SQL	false

Query

```sql
SELECT
  learner_id,
  COUNT(learning_id) AS cnt_learning,

  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q1 2018' THEN 1 ELSE 0 END) AS learning_2018Q1,
  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q2 2018' THEN 1 ELSE 0 END) AS learning_2018Q2,
  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q3 2018' THEN 1 ELSE 0 END) AS learning_2018Q3,
  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q4 2018' THEN 1 ELSE 0 END) AS learning_2018Q4,

  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q1 2019' THEN 1 ELSE 0 END) AS learning_2019Q1,
  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q2 2019' THEN 1 ELSE 0 END) AS learning_2019Q2,
  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q3 2019' THEN 1 ELSE 0 END) AS learning_2019Q3,
  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q4 2019' THEN 1 ELSE 0 END) AS learning_2019Q4,

  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q1 2020' THEN 1 ELSE 0 END) AS learning_2020Q1,
  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q2 2020' THEN 1 ELSE 0 END) AS learning_2020Q2,
  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q3 2020' THEN 1 ELSE 0 END) AS learning_2020Q3,
  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q4 2020' THEN 1 ELSE 0 END) AS learning_2020Q4,

  MAX(CASE WHEN dd.Calendar_Quarter_Year = 'Q1 2021' THEN 1 ELSE 0 END) AS learning_2021Q1,

FROM
  `scs-3760-hkim.term_project.ms_dynamics365` t1
  left join `scs-3760-hkim.term_project.dim_date` dd on t1.learning_date = dd.Date
WHERE
  EXTRACT(YEAR FROM learning_date) IN (2018,2019,2020,2021)
GROUP BY
  learner_id
```

BigQuery: Transformation

Creating a view and preparing dataset for a machine learning model

View name: v_dim_learner

Source tables

- Salesforce_contact
- V_fact_learning

View info

View ID	scs-3760-hkim:term_project.v_dim_learner
Created	22 Apr 2021, 05:18:28
Last modified	22 Apr 2021, 05:18:29
View expiry	Never
Use Legacy SQL	false

Query

```
SELECT
  t1.*,

  coalesce(t2.cnt_learning,0) AS cnt_learning,

  coalesce(t2.learning_2018Q1,0) AS learning_2018Q1,
  coalesce(t2.learning_2018Q2,0) AS learning_2018Q2,
  coalesce(t2.learning_2018Q3,0) AS learning_2018Q3,
  coalesce(t2.learning_2018Q4,0) AS learning_2018Q4,

  coalesce(t2.learning_2019Q1,0) AS learning_2019Q1,
  coalesce(t2.learning_2019Q2,0) AS learning_2019Q2,
  coalesce(t2.learning_2019Q3,0) AS learning_2019Q3,
  coalesce(t2.learning_2019Q4,0) AS learning_2019Q4,

  coalesce(t2.learning_2020Q1,0) AS learning_2020Q1,
  coalesce(t2.learning_2020Q2,0) AS learning_2020Q2,
  coalesce(t2.learning_2020Q3,0) AS learning_2020Q3,
  coalesce(t2.learning_2020Q4,0) AS learning_2020Q4,

  coalesce(t2.learning_2021Q1,0) AS learning_2021Q1

FROM
  `scs-3760-hkim.term_project.salesforce_contact` t1
LEFT JOIN
  `scs-3760-hkim.term_project.v_fact_learning` t2
ON
  t1.learner_id = t2.learner_id
```

AI Platform:
Jupyter Notebook

Part 1 - Load dataset from BigQuery

- Source table: v_dim_learner
- Target dataframe: loaded

Part 2 - Prep and train a ML model

- Binary classification model
  - 1 = take ALR courses
  - 0 = not take ALR courses
- Prediction dataframe: df_Y_pred

Part 3 - Load prediction to BigQuery

- Source dataframe: df_Y_pred
- Target table: learning_pred

Part 1

```
# ========================
# 1. Setup & Load dataset
# ========================
# a) Load libraries
# python libs
import numpy as np
import pandas as pd
# import matplotlib.pyplot as plt
from pandas import read_csv
from pandas import set_option

# data transformation
from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import LabelEncoder

# classification models
from sklearn.ensemble import GradientBoostingClassifier

# b) Load dataset from bigqquery table
from google.cloud import bigquery
client = bigquery.Client(location="US")
query = "SELECT * FROM `scs-3760-hkim.term_project.v_dim_learner`"
query_job = client.query(query, location="US")
loaded = query_job.to_dataframe()
```

Part 2

```
'''
Train a machine learning model and make predictions
'''
```

Part 3

```
# ================================
# 8. Load Prediction to bigquery
# ================================
# drop table learning_pred
sql = "drop table if exists `scs-3760-hkim.term_project.learning_pred`"
client.query(sql, location="US")

# load
bg_dataset = client.dataset('term_project')
table_ref = bg_dataset.table("learning_pred")
job = client.load_table_from_dataframe(df_Y_pred, table_ref, location="US")

job.result()  # Waits for table load to complete.
print("Loaded dataframe to {}".format(table_ref.path))
```

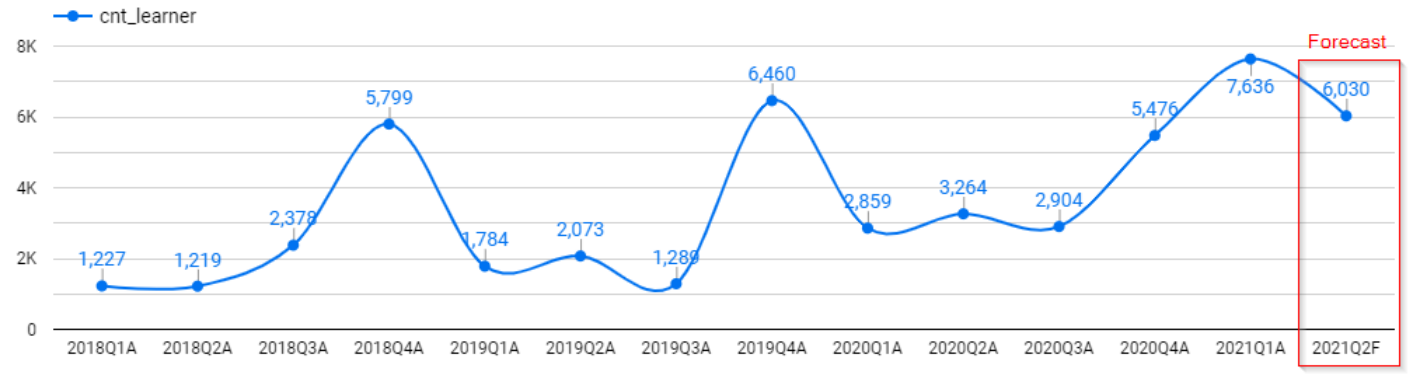
Data Studio: Visualization

Data source: v_fact_learning_sum_pred

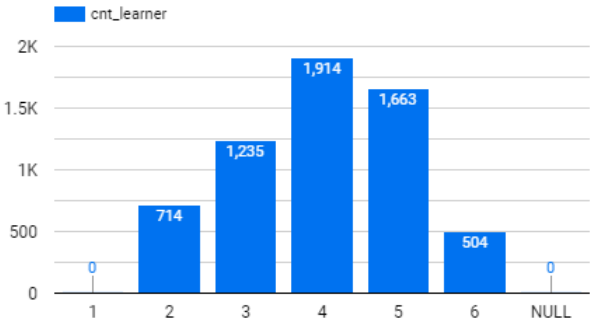
When the report was pointed to external tables (in Cloud Storage), it has some performance issues.

Instead of external tables, created BigQuery tables with AI Platform Notebooks (01-load_cs_to_bq.ipynb).

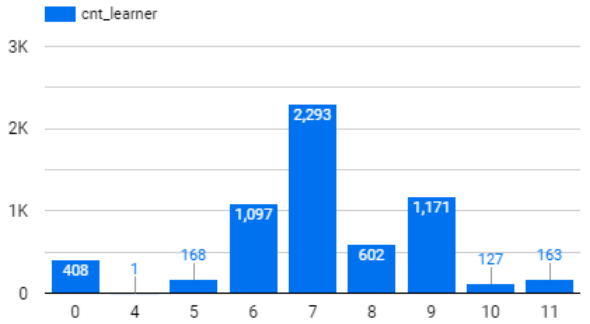
How are numbers of learners trending?



2021Q2F - Forecast by Age Group



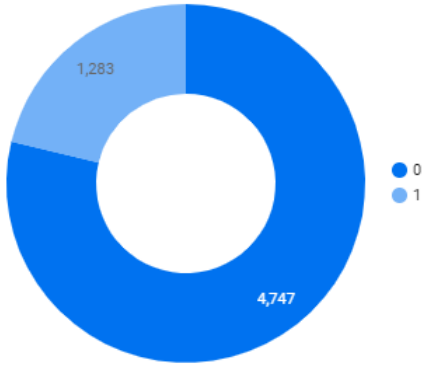
2021Q2F - Forecast by Avg. Income Group



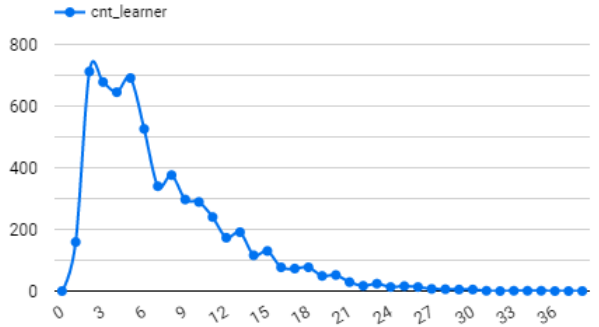
Data Studio: Visualization

Data source: v_fact_learning_sum_pred

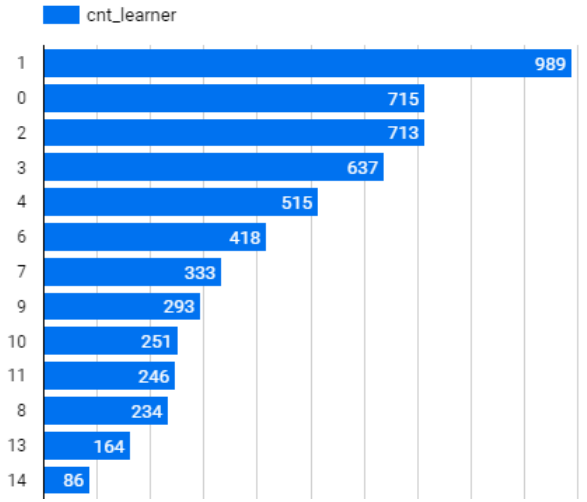
2021Q2F - Forecast by C-Level



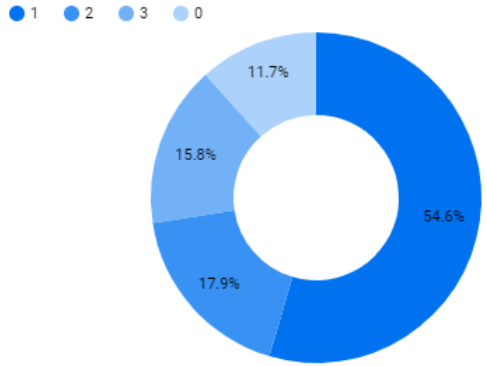
2021Q2F - Forecast by # of Learning (last 3 years)



2021Q2F - Forecast by Industry



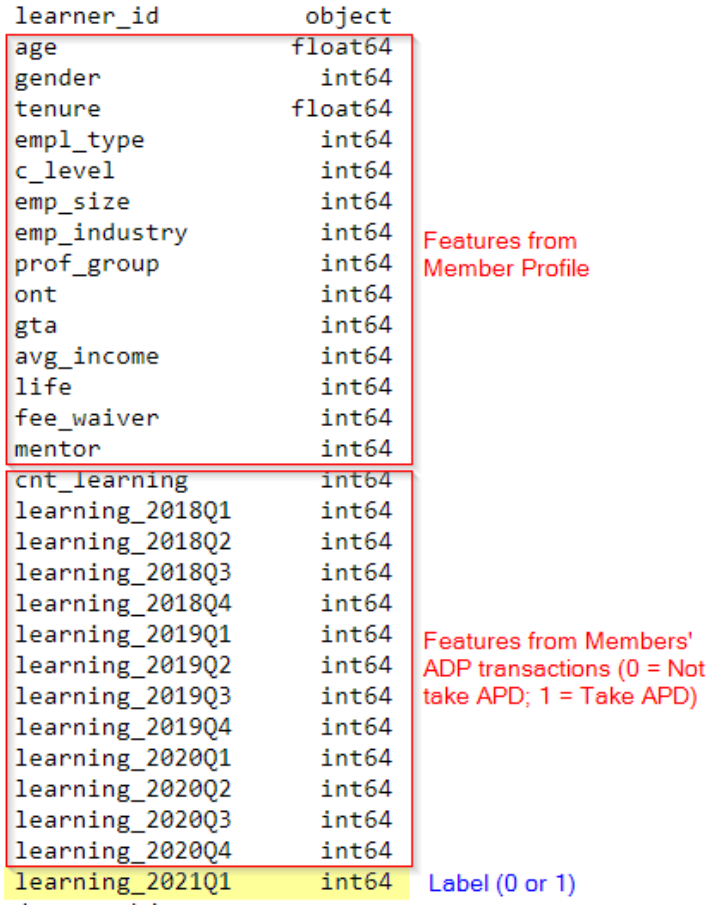
2021Q2F - Forecast by Employer Size



Appendix

Appendix A – Label and Features

- Label
  - 1 = Take ALR
  - 0 = Not take ALR
- Features
  - Member profile
  - Member ALR transactions (last 3-years)



Appendix A – Gartner Magic Quadrant for Cloud infrastructure and Platform Services

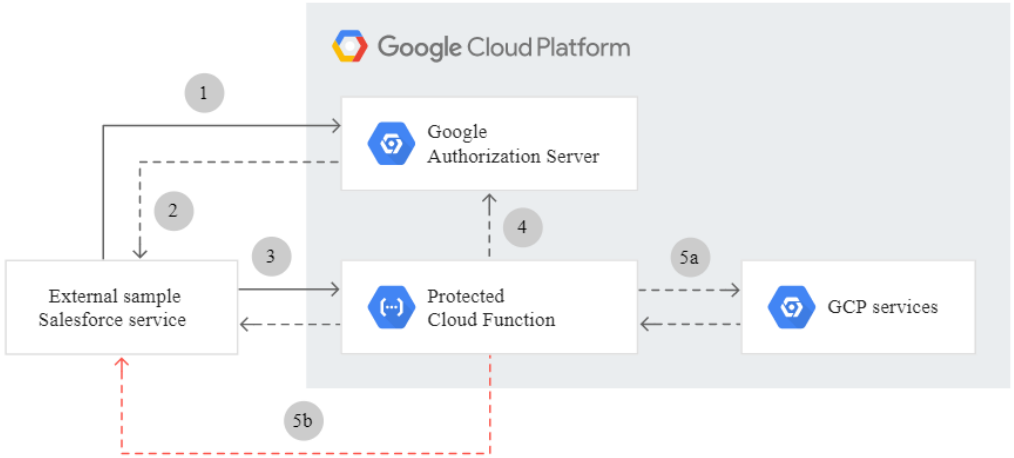
Figure 1. Magic Quadrant for Cloud Infrastructure and Platform Services

The capability gap between hyperscale cloud providers has begun to narrow; however, fierce competition for enterprise workloads extends to secondary markets worldwide. Infrastructure and operations leaders should evaluate cloud providers with a broad range of use cases and a wide market presence.

https://www.gartner.com/en/documents/3989743/magic-quadrant-for-cloud-infrastructure-and-platform-ser

Appendix C – Data Integration with Salesforce

Integrating Salesforce CRM with Cloud Functions



You can integrate Cloud Functions into your ecosystem, where they can serve as a building block in your end-to-end enterprise business process. You can use Cloud Functions for tasks such as the following:

- Inserting business data into Google Cloud for storage and analytical processing, such as customer records from front-end CRM systems.
- Creating or updating customer data held in backing stores such as Firestore and Cloud SQL when those stores serve as customer masters (that is, as the system of record).
- Retrieving transactional data such as orders, service requests, service appointments, and product details from data stores on Google Cloud. You might do this in order to create customer 360-degree views in an SaaS platform such as Salesforce.
- Transforming data files received from partner organizations that need to be parsed, processed, and then loaded into a data lake or data warehouse on Google Cloud.
- Parsing consumer interaction data such as form submissions or image or document uploads to a website for generating insights using BigQuery and BigQuery ML.

https://cloud.google.com/architecture/calling-protected-cloud-functions

Appendix D BigQuery: Transformation

View name: v_dim_learner_pred

Data source

- v_dim_learner
- learning_pred (loaded from AI Platforms)

Appendix E BigQuery: Transformation

Prepare dataset for visualization and reports

Converted to transactional dataset

View name: v_fact_learning_sum_pred

Source

- Tables: v_dim_learner_pred

v_fact_learning_sum_pred

QUERY VIEW SHARE VIEW COPY VIEW

View info

View ID	scs-3760-hkim:term_project.v_fact_learning_sum_pred
Created	22 Apr 2021, 07:14:08
Last modified	22 Apr 2021, 09:54:06
View expiry	Never
Use Legacy SQL	false

Query

```
SELECT '2018Q1A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2018Q1) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2018Q2A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2018Q2) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2018Q3A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2018Q3) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2018Q4A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2018Q4) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL

SELECT '2019Q1A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2019Q1) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2019Q2A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2019Q2) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2019Q3A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2019Q3) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2019Q4A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2019Q4) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL

SELECT '2020Q1A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2020Q1) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2020Q2A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2020Q2) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2020Q3A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2020Q3) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2020Q4A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2020Q4) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL

SELECT '2021Q1A' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2021Q1) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 UNION ALL
SELECT '2021Q2F' AS yr_qtr,age,gender,tenure,empl_type,c_level,emp_size,emp_industry,prof_group,ont,gta,avg_income,life,fee_waiver,mentor,cnt_learning,
SUM(learning_2021Q2_pred) AS cnt_learner FROM `scs-3760-hkim.term_project.v_dim_learner_pred` GROUP BY 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16
```