Building ML End-to-End

Google Cloud Platform

Presenter: Federico Arroyo

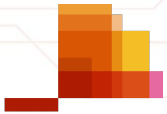
Federico Arroyo

- 6 years in data science consulting
- Data Science Manager @ Alvarez & Marsal
- Technical on-the-ground expertise
  - Cloud-based ML (AWS, Azure, GCP)
  - Time-Series Modelling
  - Natural Language Processing
  - MLOps implementation

Coming Soon

Professional Experience

Companies



pwc

Capgemini

Commercial Clients

KAISER PERMANENTE®

GUIDEWELL

Lilly

Pfizer

WARNER BROS.

ROGERS™

DIAGEO

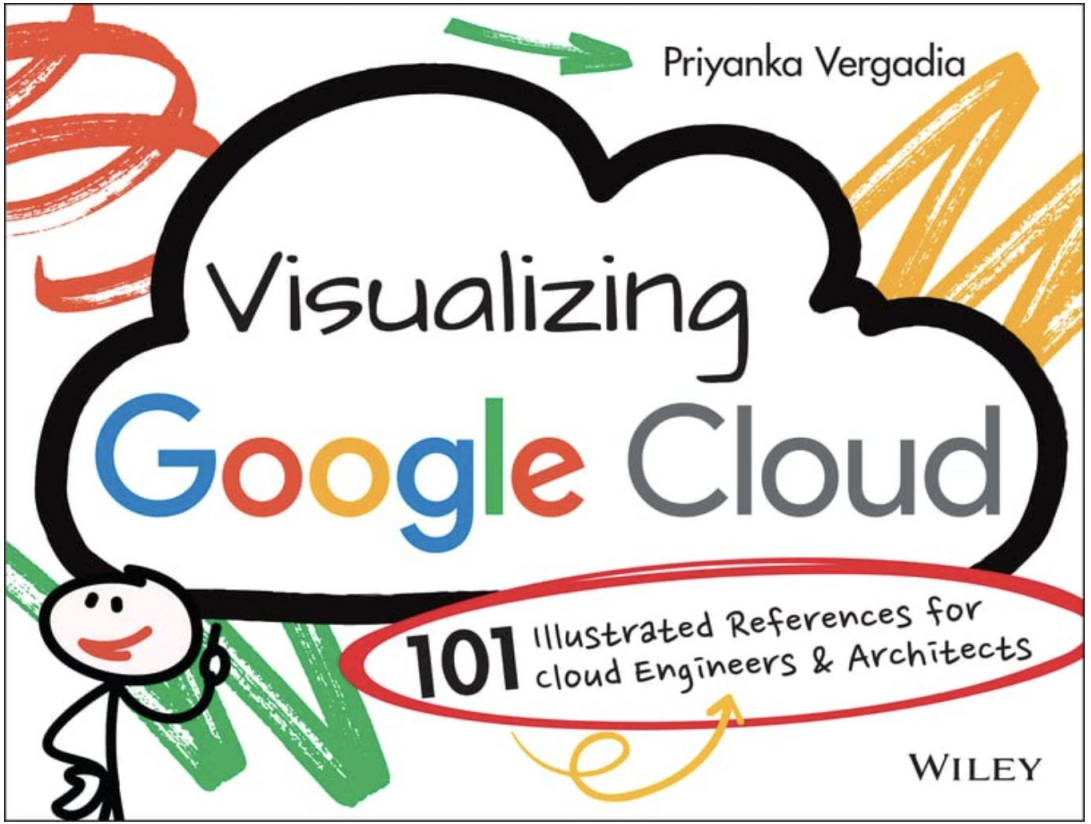
Learning Objectives for Today

1. Learn about Cloud History and its practical use cases for Machine Learning and AI applications

2. Understand the Machine Learning Lifecycle as it pertains to google cloud platform

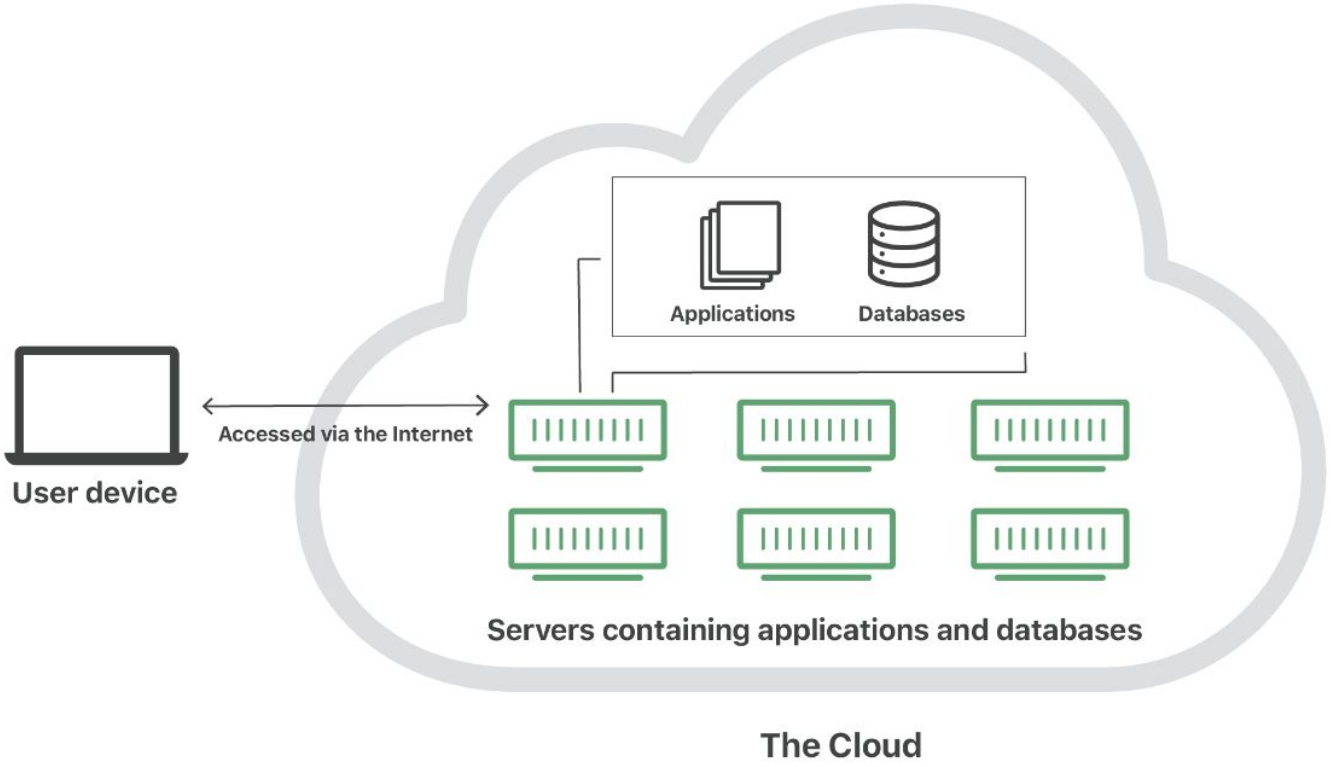
3. Apply understanding in a live demo detecting fraud for a large financial institution using Machine Learning

Want to learn more?



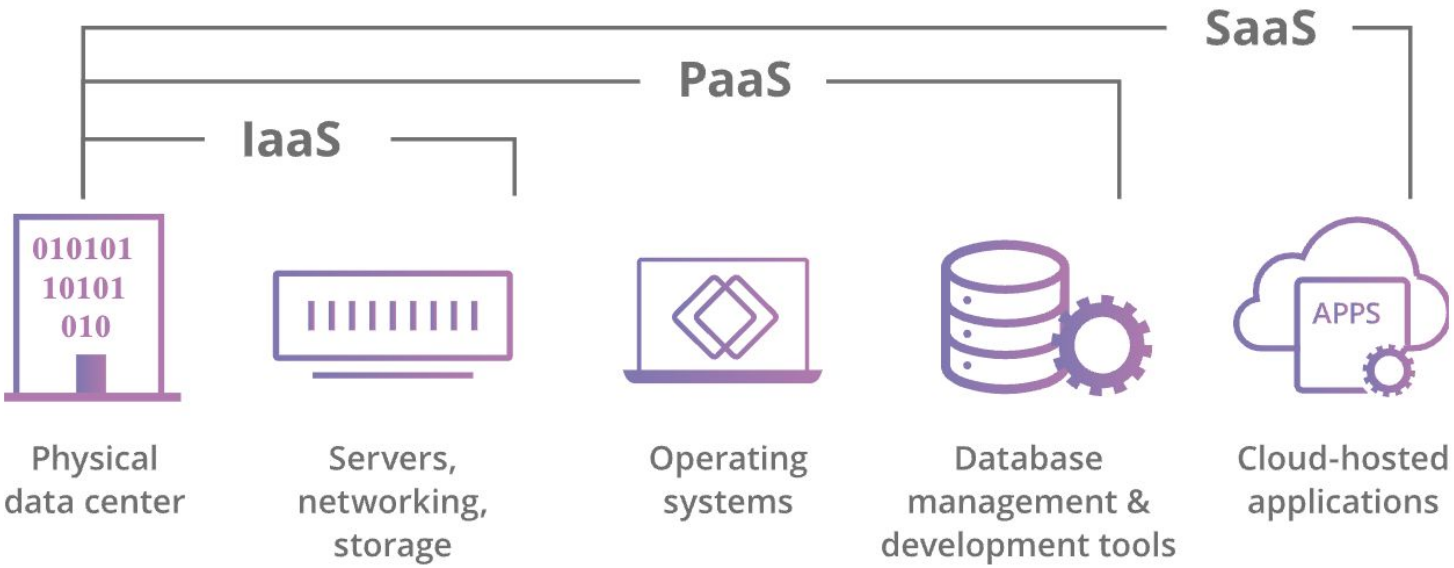

My boss told me to put my files in the cloud.
TheLifeOfDanger.com

What is the cloud?



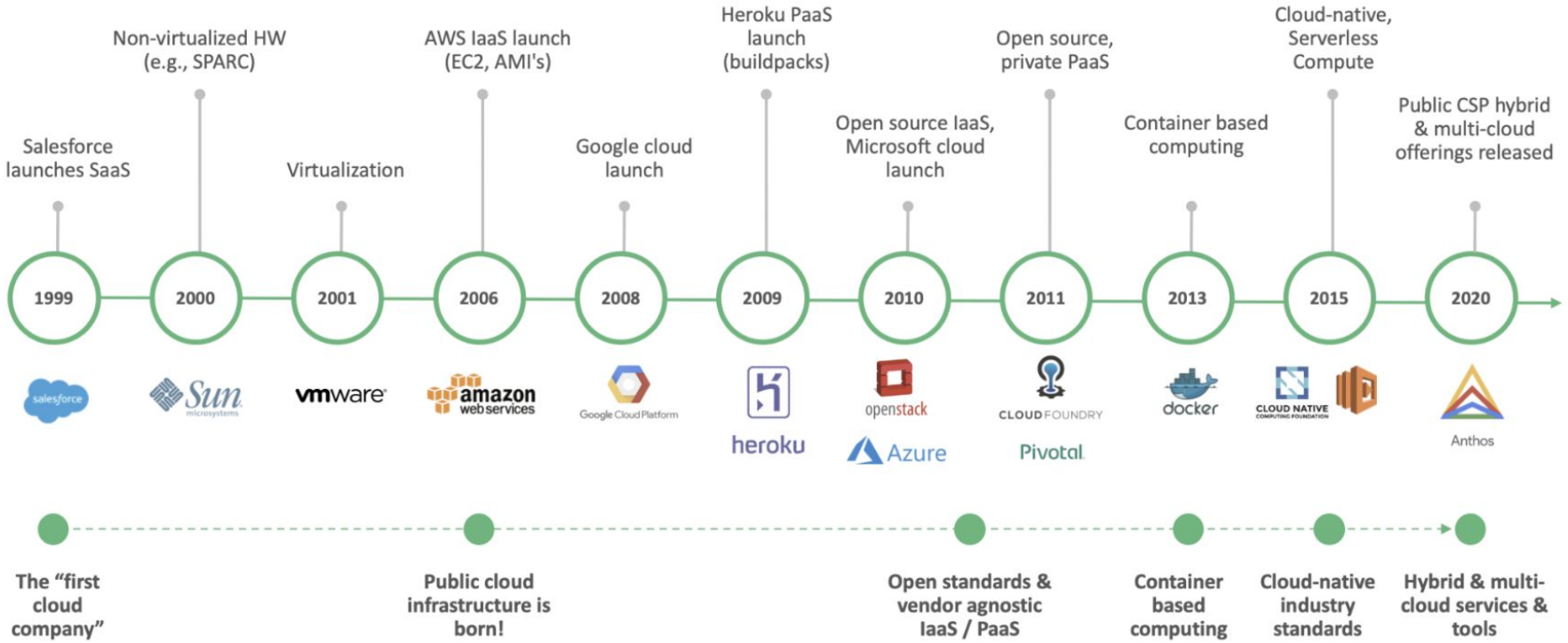
Source: https://www.cloudflare.com/learning/cloud/what-is-the-cloud/

What are the main service models?



Source: https://www.cloudflare.com/learning/cloud/what-is-the-cloud/

History



Source: Cloud-Native Computing Foundation: https://www.cncf.io/blog/2017/05/15/developing-cloud-native-applications/| Cloud Native and Container Technology Landscape

Service comparison

LINK

	SERVICE TYPE	DESCRIPTION	aws	Azure	Google Cloud
STORAGE	Object storage	For storing any files you regularly use	Simple Storage Service (S3)	Blob Storage	Cloud Storage Buckets
	Archive storage	Low cost (but slower) storage for rarely used files	S3 Glacier Instant, Glacier Flexible, Glacier	Blob Cool/Cold/Archive tiers	Cloud Storage Nearline, Coldline, Archive
	File storage	For storing files needing hierarchical organization	Elastic File System (EFS), FSx	Avers vFXT, Files	Filestore
	Block storage	For storing groups of related files	Elastic Block Storage	Disk Storage	Persistent Disk
	Hybrid storage	Move files between on-prem & cloud	Storage Gateway	StorSimple, Migrate	Storage Transfer Service
	Edge/offline storage	Move offline data to the cloud	Snowball	Data Box	Transfer Appliance
	Backup	Prevent data loss	Backup	Backup	Backup and Disaster Recovery
DATABASE	Relational DB management	Standard SQL DB (PostgreSQL, MySQL, SQL Server, etc.)	Relational Database Service (RDS), Aurora	SQL, SQL Database	Cloud SQL, Cloud Spanner
	NoSQL: Key-value	Redis-like DBs for semi-structured data	DynamoDB	Cosmos DB, Table storage	Cloud BigTable, Firestore
	NoSQL: Document	MongoDB/CouchDB-like DBs for hierarchical JSON data	DocumentDB	Cosmos DB	Firestore, Firebase Realtime Database
	NoSQL: Column store	Cassandra/HBase-like DBs for structured hierarchical data	Keyspaces	Cosmos DB	Cloud BigTable
	NoSQL: Graph	Neo4j-like DBs for connected data	Neptune	N/A	N/A
	Caching	Redis/Memcached-like memory for calculation	ElastiCache	Cache for Redis, HPC Cache	Memorystore
	Time Series DB	DB tuned for time series data	Timestream	Time Series Insights	Cloud BigTable
	Blockchain	Dogecoin, etc.	Managed Blockchain	Blockchain Service, Blockchain Workbench, Confidential Ledger	N/A

Source: DataCamp

Questions about Cloud?

Let’s jump into GCP!

Google Cloud Platform

How to get started?

90-day, $300 Free Trial: New Google Cloud and Google Maps Platform users can take advantage of a 90-day trial period that includes $300 in free Cloud Billing credits to explore and evaluate Google Cloud and Google Maps Platform products and services. You can use these credits toward one or a combination of products.

What do I get?

If you upgrade before the trial is over: Any remaining, unexpired Free Trial credits remain in your Cloud Billing account. You can continue to use the resources you created during the Free Trial without interruption.

For resources you use in excess of what's covered by any remaining credit, your form of payment on file is charged (credit card or bank account).

If you upgrade within 30 days of the end of the trial: Your resources are marked for deletion, but you might be able to recover them. Learn more about data deletion on Google Cloud.

If you upgrade more than 30 days after the end of the trial, your Free Trial resources are lost.

How to get started?

Check it out

90-day, $300 Free Trial: New Google Cloud and Google Maps Platform users can take advantage of a 90-day trial period that includes $300 in free Cloud Billing credits to explore and evaluate Google Cloud and Google Maps Platform products and services. You can use these credits toward one or a combination of products.

$300

90 days

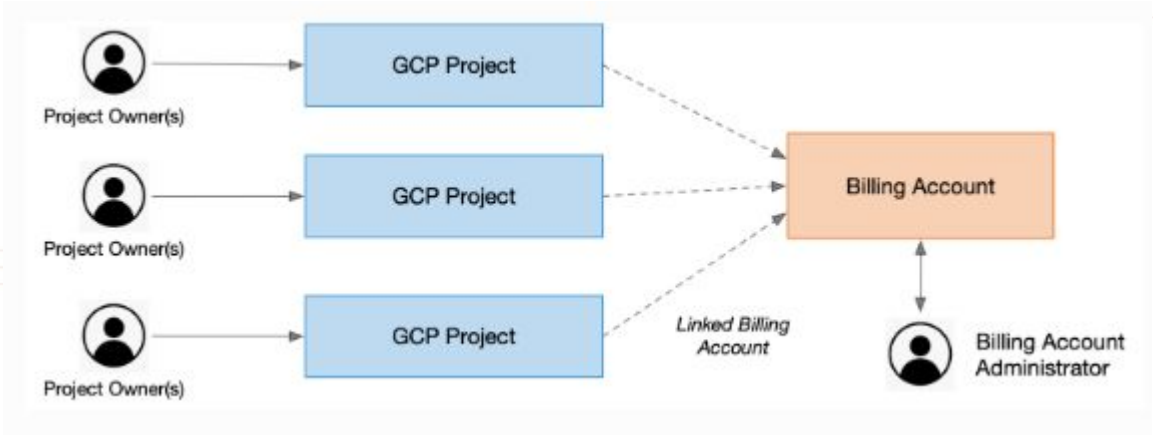
If you upgrade before the trial is over: Any remaining, unexpired Free Trial credits remain in your Cloud Billing account. You can continue to use the resources you created during the Free Trial without interruption.

For resources you use in excess of what's covered by any remaining credit, your form of payment on file is charged (credit card or bank account).

If you upgrade within 30 days of the end of the trial: Your resources are marked for deletion, but you might be able to recover them. Learn more about data deletion on Google Cloud.

If you upgrade more than 30 days after the end of the trial, your Free Trial resources are lost.

Billing



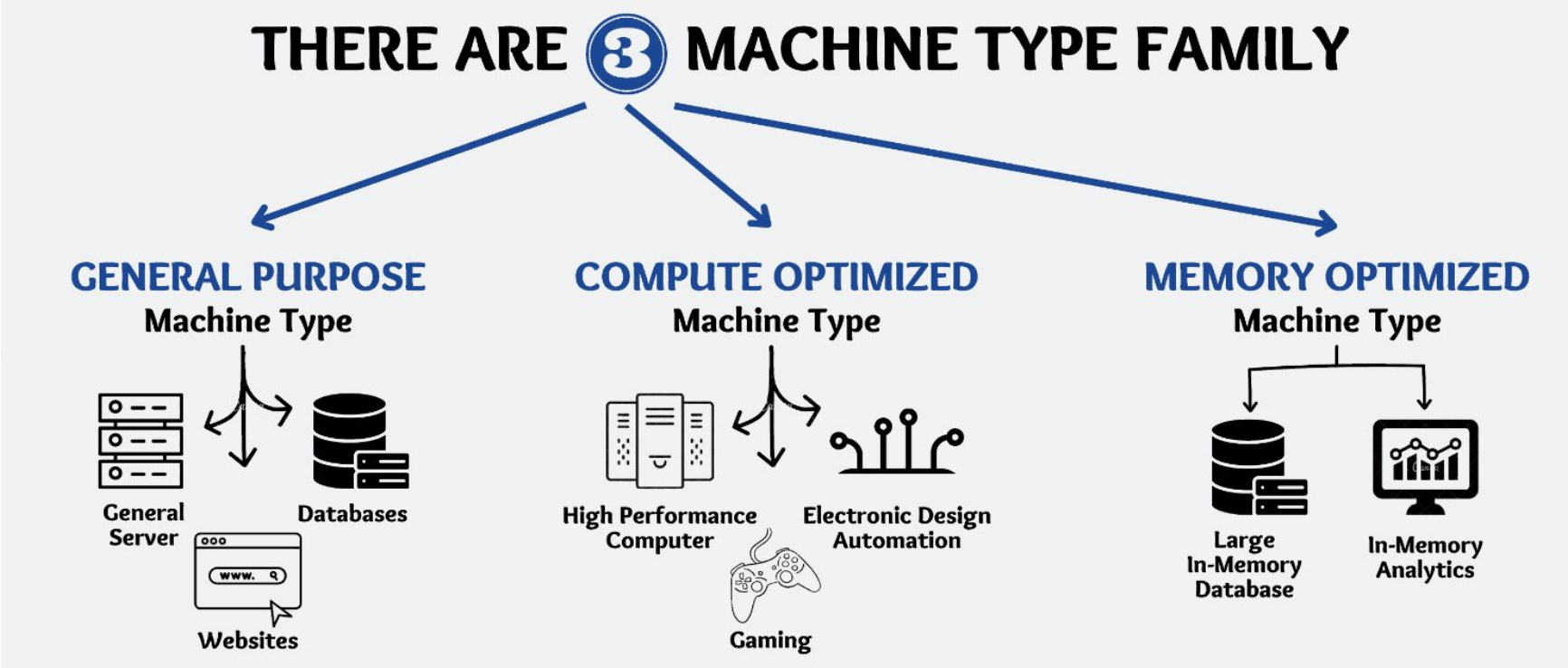
Source: https://docs.dimensions.ai/bigquery/gcp-setup.html

Compute

CATEGORY	PURPOSE
Standard	Balanced between processing power and memory. Fits most common application needs
High-Memory	Emphasis is put on memory over processing power for tasks that need accessible non-disk storage quickly
High-CPU	Higher CPU usage for high-intensity applications that require processing over memory
Shared-core	A single virtual CPU, backed by a physical CPU, that can run for a period of time. These machines are not for use cases that require an ongoing server or significant power. The micro shared-core machine also provides bursting capability when the virtual CPU requires more power than the single physical core. Bursting is for a short, intermittent period based on need.

Source: https://k21academy.com/google-cloud/google-compute-engine/

Compute



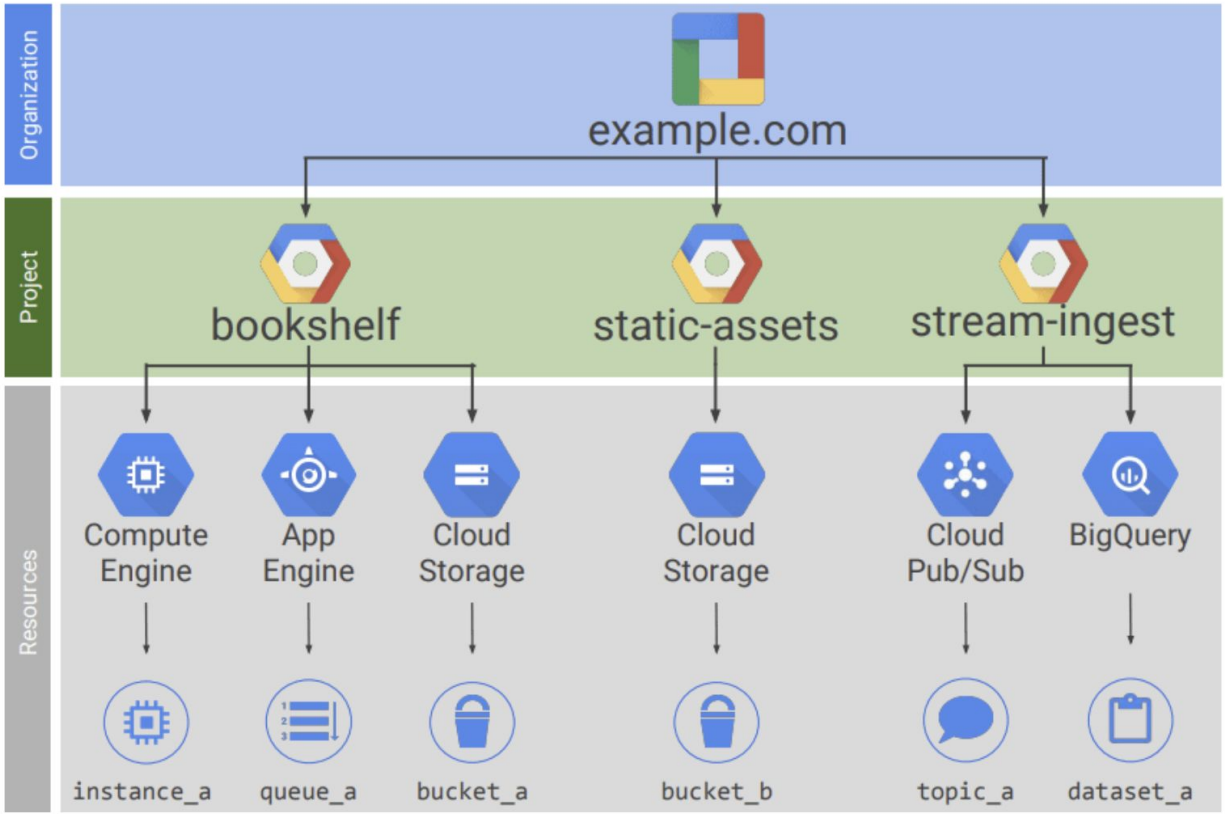
Source: https://k21academy.com/google-cloud/google-compute-engine/

Storage

	Cloud Datastore	Bigtable	Cloud Storage	Cloud SQL	Cloud Spanner	BigQuery
Type	NoSQL document	NoSQL Wide Column	Blob Storage	Relational SQL - OLTP	Relational SQL - OLTP	Relational SQL - OLAP
Transactions	Yes	Single-Row	No	Yes	Yes	No
Complex Queries	No	No	No	Yes	Yes	Yes
Capacity	Terabytes	Petabytes	Petabytes	Upto ~10TB	Petabytes	Petabytes
Unit Size	1MB/entity	~10 MB/cell ~100 MB/row	5TB per object	Depends on DB Engine	10,240 MiB/row	19 MB/row
Best for	Getting started, App Engine application	Flat data, heavy read/write events, analytical data	Structured or unstructured binary or object data	Web frameworks, existing applications	Large-scale database applications	Interactive querying, offline analytics

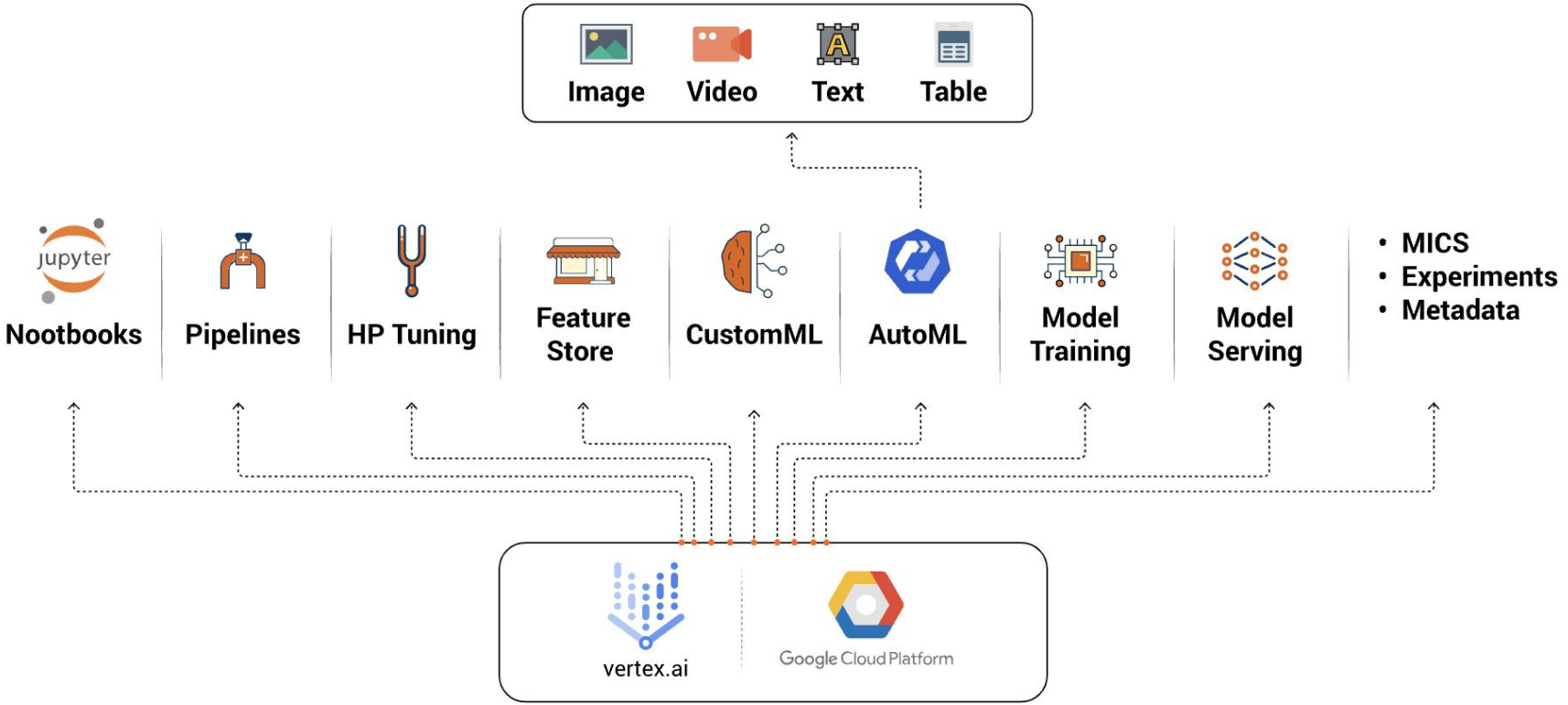
Source: https://jayendrapatil.com/google-cloud-storage-options/#google_vignette

GCP Hierarchy



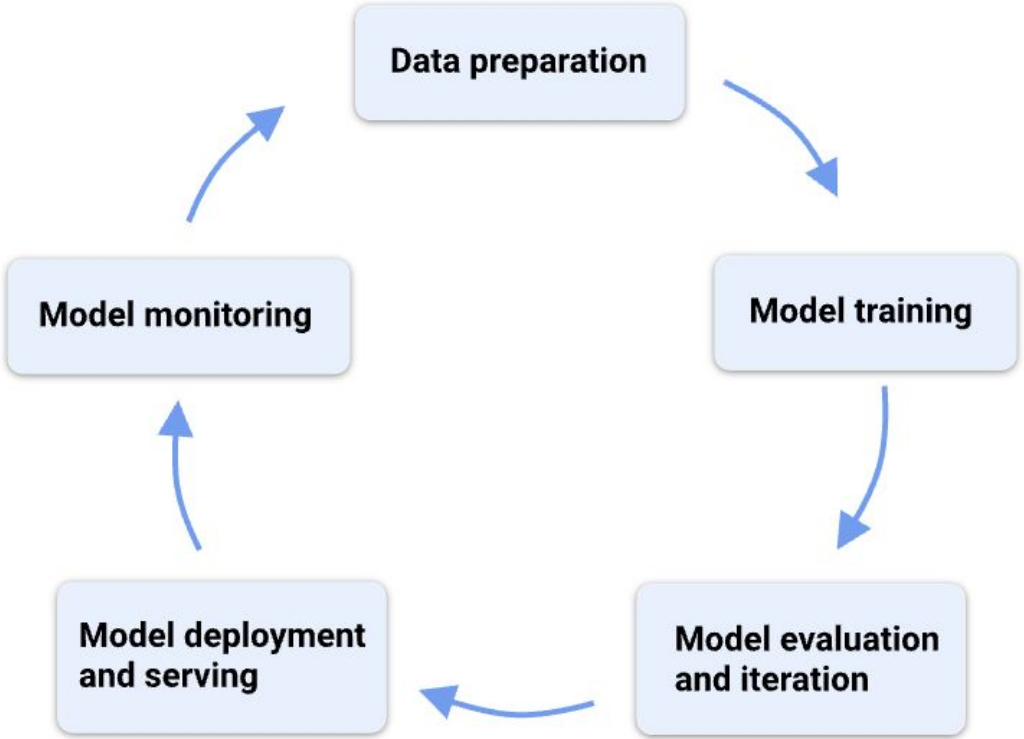
Source: https://k21academy.com/google-cloud/google-cloud-projects/

Vertex AI



Source: https://www.royalcyber.com/webinar/machine-learning-models-with-vertex-ai/

Machine learning workflow



Source: https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform

Machine Learning Lifecycle

Preprocess data → EDA → Feature Engineering → Model Selection → Model Tuning → Evaluation → Predictions

The very beginning…

Caveat

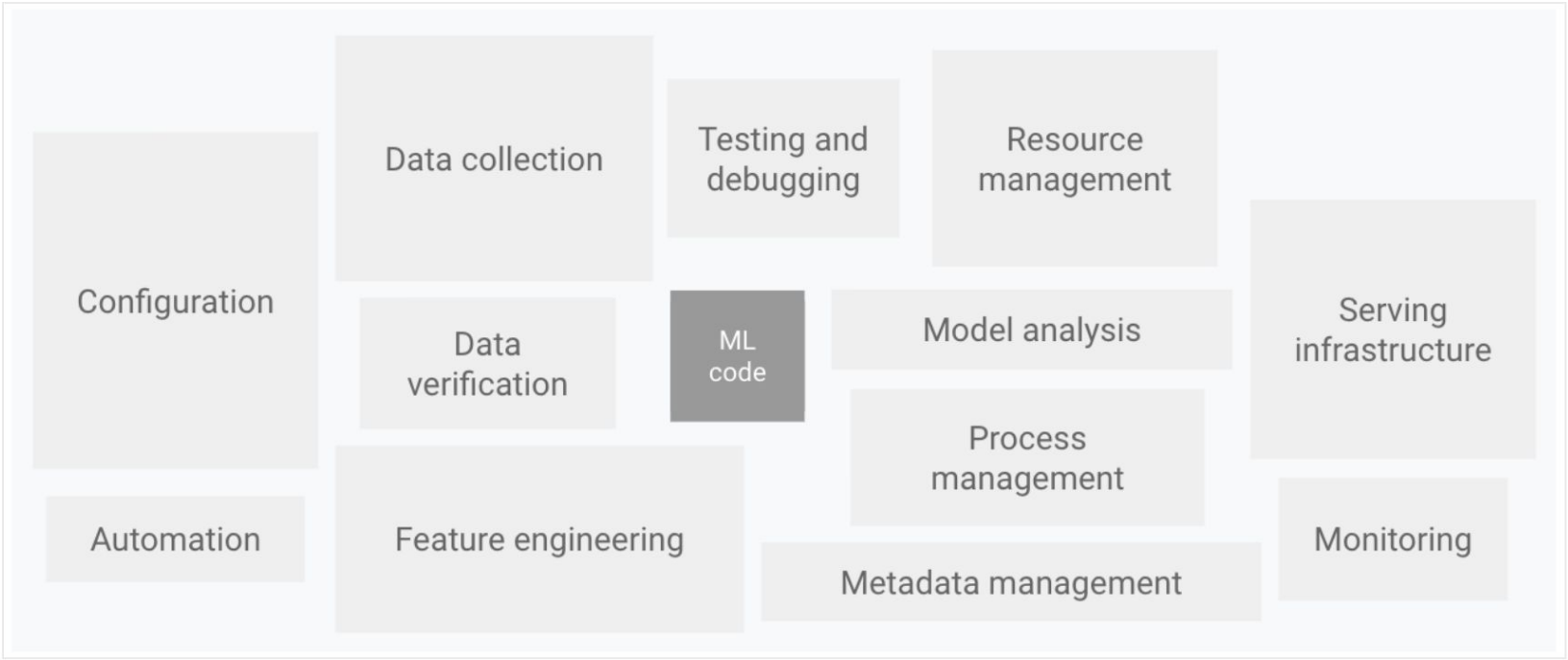


Figure 1. Elements for ML systems. Adapted from Hidden Technical Debt in Machine Learning Systems.

Machine Learning Lifecycle

Streamlit

aws

Model Number	Model Type	Time to Run (min)	Library	Hyperparameters	Features Used
1	Linear Regression	2:30	sklearn	Lasso technique	Sales Lag 1 year, Sales Lag 1 week, Sales Lag 2 weeks, Promotion flag, Weekend flag, Location Dummies
2	Decision Tree	5:00	sklearn	max_depth = 12 min_samples_split = 3 max_features = 8	Sales Lag 1 month, Sales Lag 2 weeks, Promotion flag, Weekend flag, Location Dummies
3	Decision Tree	5:30	sklearn	max_depth = 26 min_samples_split = 5 max_features = 9	Sales Lag 1 week, Sales Lag 2 weeks, Promotion flag, Weekend flag, State Dummies
4	Random Forest	7:30	sklearn	n_estimators = 12 max_depth = 9	Sales Lag 1 week, Sales Lag 2 weeks, Promotion flag, Weekend flag, ZIP Code Dummies
5	Random Forest	6:59	sklearn	n_estimators = 8 max_depth = 9	New Item Flag, Sales Lag 1 week, Sales Lag 2 weeks, Promotion flag, Weekend flag, Location Dummies
6	XGBoost	8:34	xgboost	eta = 0.4 max_depth = 9 sampling_method = uniform	Sales Lag 1 week, Sales Lag 2 weeks, Promotion flag, Weekend flag, Location Dummies
7	XGBoost	7:45	xgboost	eta = 0.2 max_depth = 5 sampling_method = uniform	Price, Sales Lag 1 week, Sales Lag 2 weeks, Promotion flag, Weekend flag, Location Dummies

Predictions

Preprocess data

EDA

Feature Engineering

Model Selection

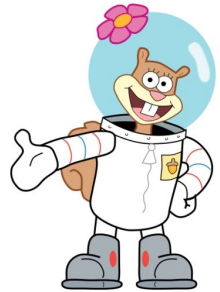
Model Tuning

Evaluation

.txt

.txt

Technical Doc Template



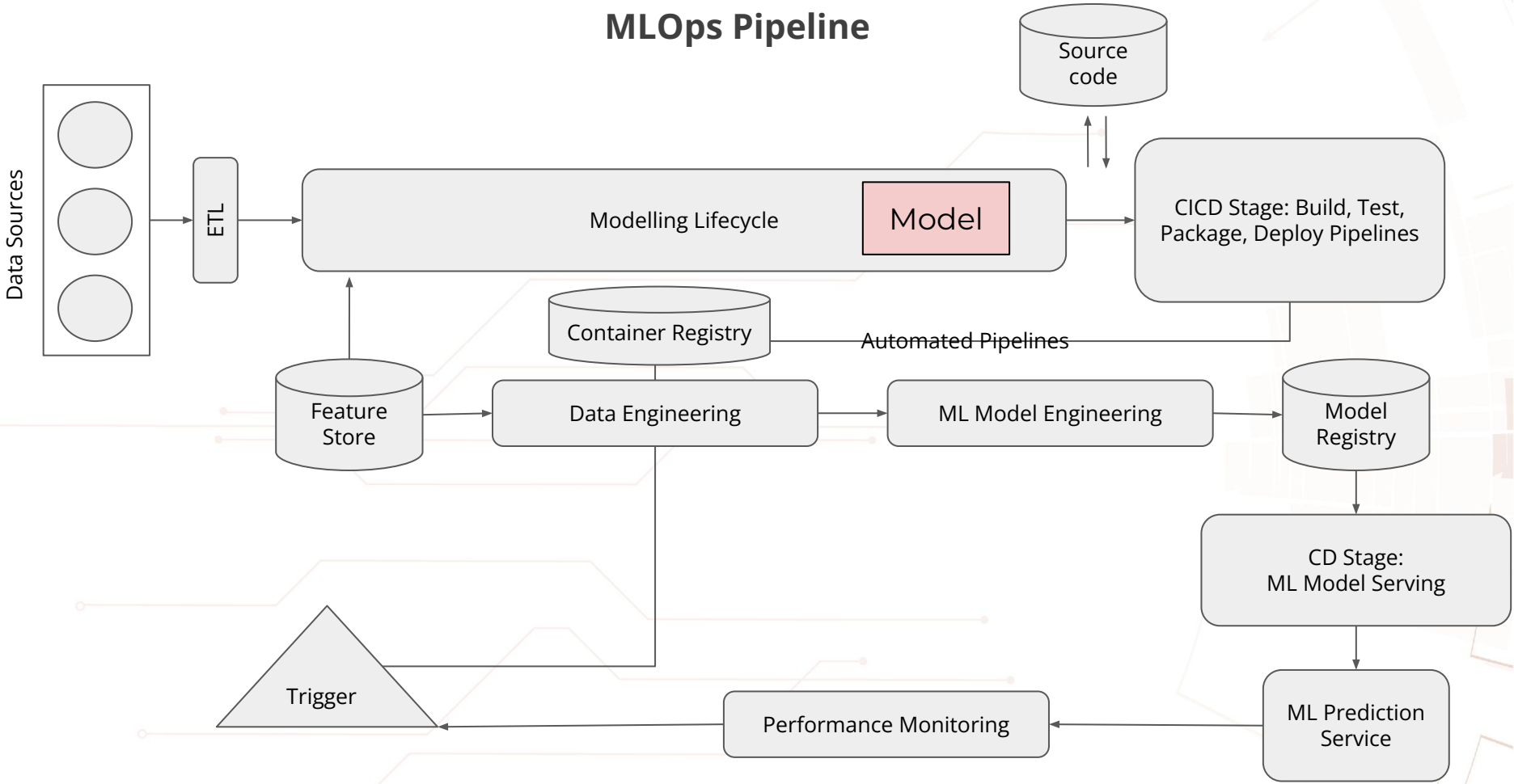
Powered by simplilearn

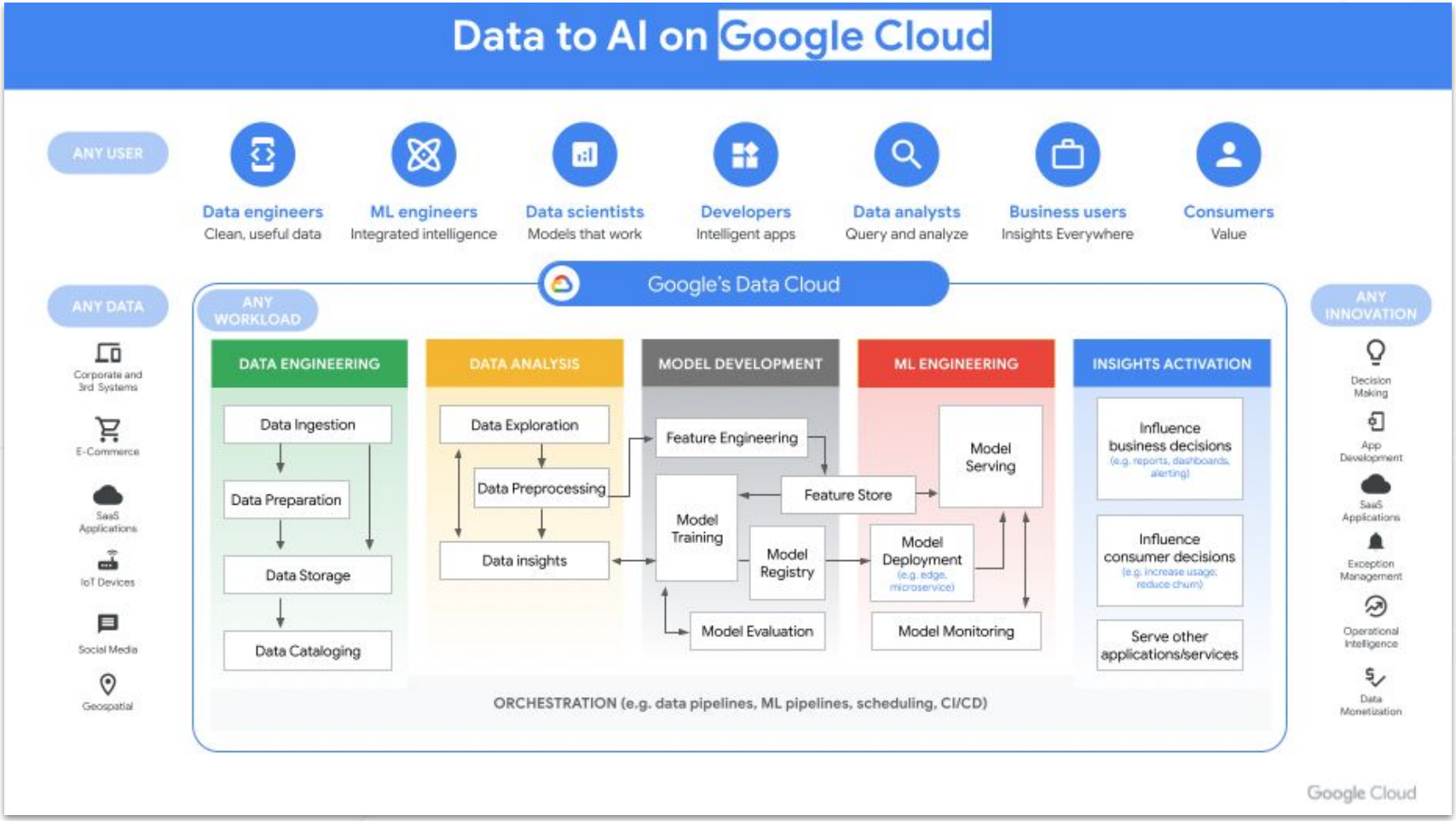
Caltech | Center for Technology & Management Education

ML lifecycle that only focuses on modelling

Source: Completeset.com

MLOps Pipeline




Data to AI on Google Cloud
ANY USER
Data engineers
Clean, useful data
ML engineers
Integrated intelligence
Data scientists
Models that work
Developers
Intelligent apps
Data analysts
Query and analyze
Business users
Insights Everywhere
Consumers
Value
Google's Data Cloud
ANY DATA
Corporate and 3rd Systems
E-Commerce
SaaS Applications
IoT Devices
Social Media
Geospatial
ANY WORKLOAD
DATA ENGINEERING
Data Ingestion
Data Preparation
Data Storage
Data Cataloging
DATA ANALYSIS
Data Exploration
Data Preprocessing
Data insights
MODEL DEVELOPMENT
Feature Engineering
Model Training
Model Registry
Model Evaluation
ML ENGINEERING
Feature Store
Model Serving
Model Deployment
(e.g. edge, microservice)
Model Monitoring
INSIGHTS ACTIVATION
Influence business decisions
(e.g. reports, dashboards, alerting)
Influence consumer decisions
(e.g. increase usage, reduce churn)
Serve other applications/services
ORCHESTRATION (e.g. data pipelines, ML pipelines, scheduling, CI/CD)
ANY INNOVATION
Decision Making
App Development
SaaS Applications
Exception Management
Operational Intelligence
Data Monetization
Google Cloud

LINK

Want to learn more?

StatMike

@statmike-channel · 1.8K subscribers · 12 videos

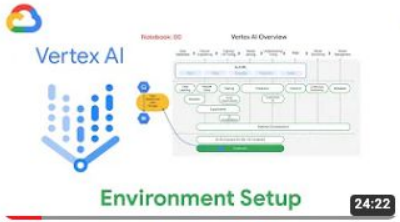
Hi, I'm Mike 👋 >

github.com/statmike and 1 more link

Subscribed

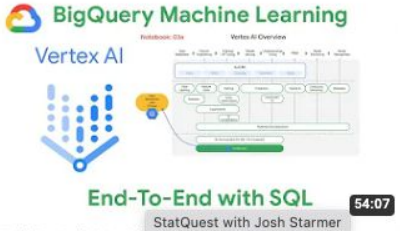
Home Videos Playlists Community

For You



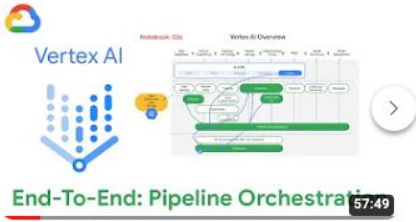
Environment Setup - Vertex AI for ML Operations [notebook 00]

14K views • 1 year ago



End-To-End: ML with SQL in BigQuery (BQML) [notebook 03a]

3.5K views • 1 year ago

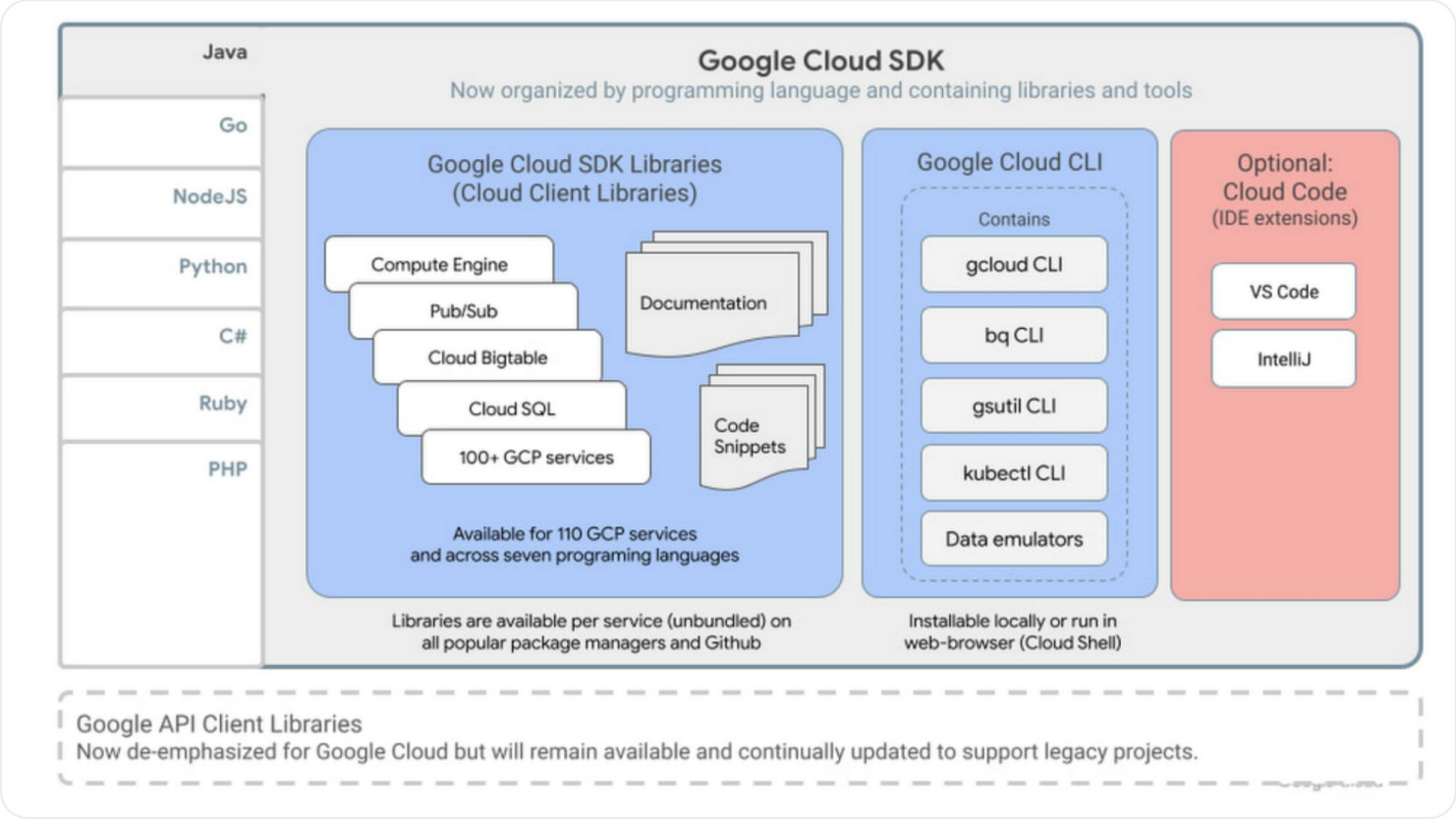


End-To-End: Pipeline Orchestration (KFP) - AutoML in Vertex AI for ML Operations [notebook 02c]

6.6K views • 1 year ago

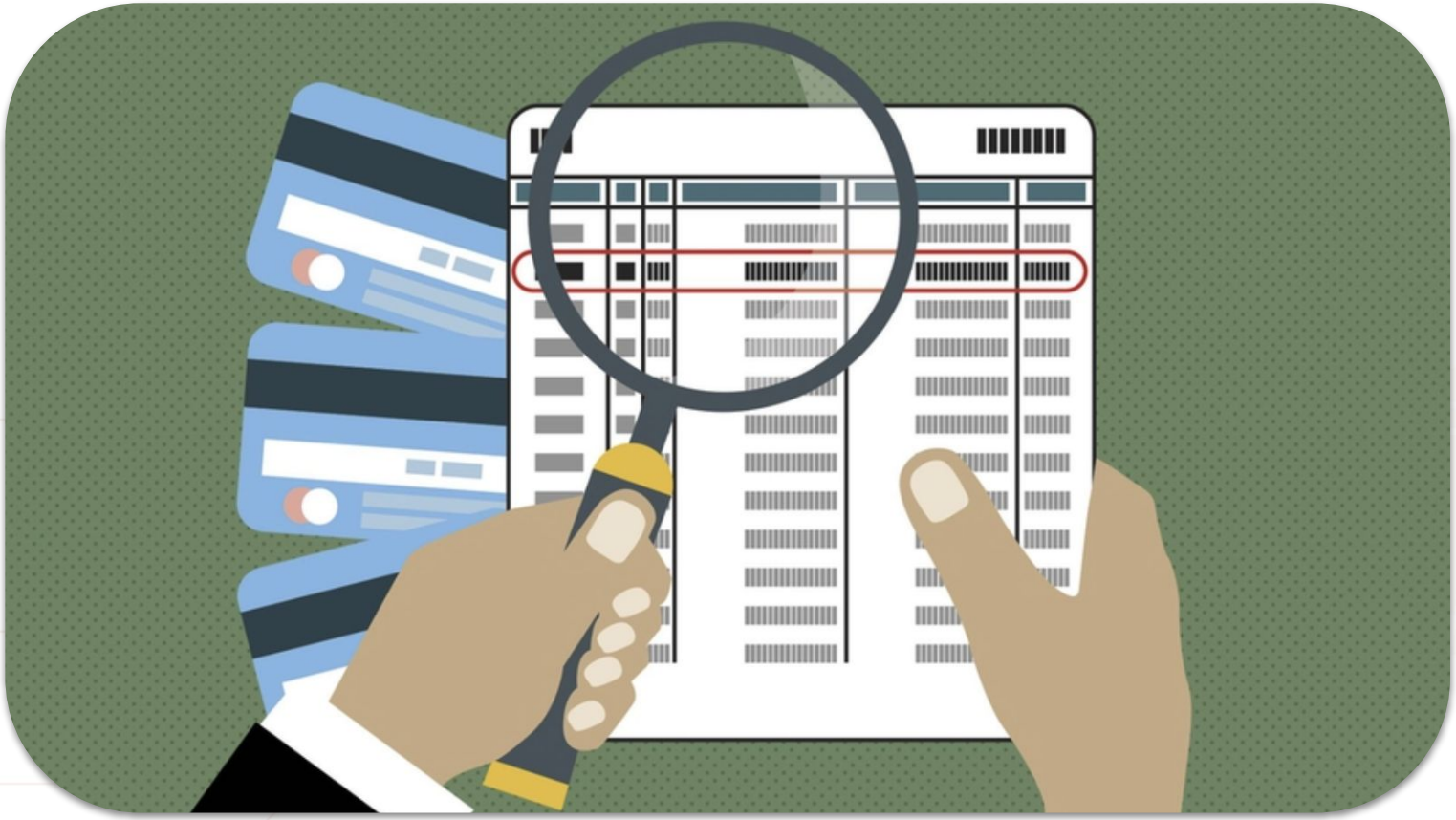
Cloud SDK

LINK (how to)



Source: Google documentation

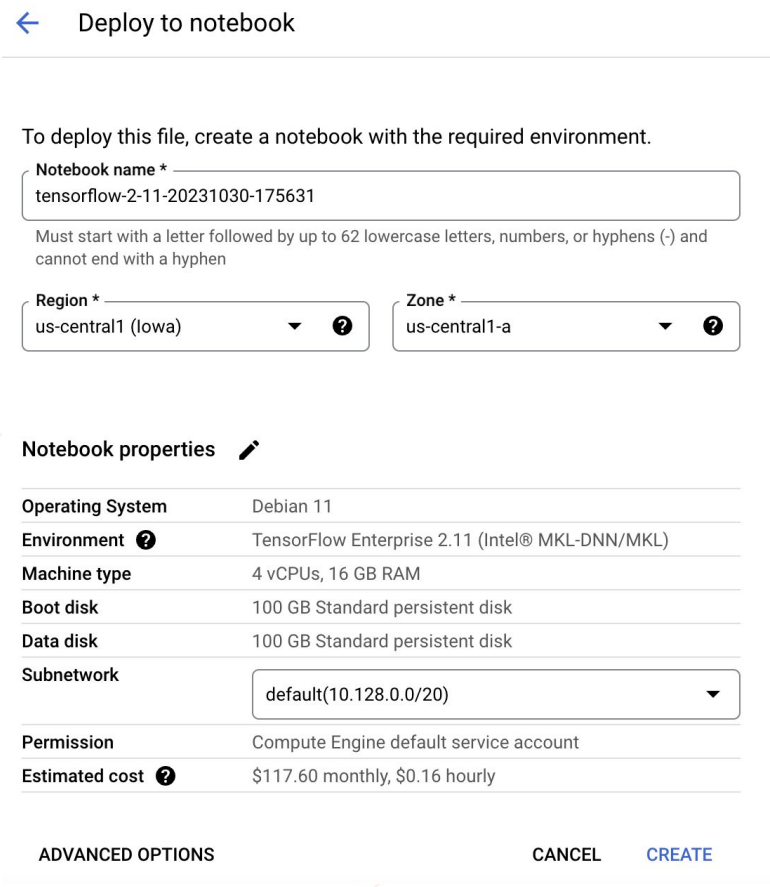
Today’s example



Source: MIT News

Spinning up Workbench (notebook)

Need a guide?



What is Tensorflow?

Regions

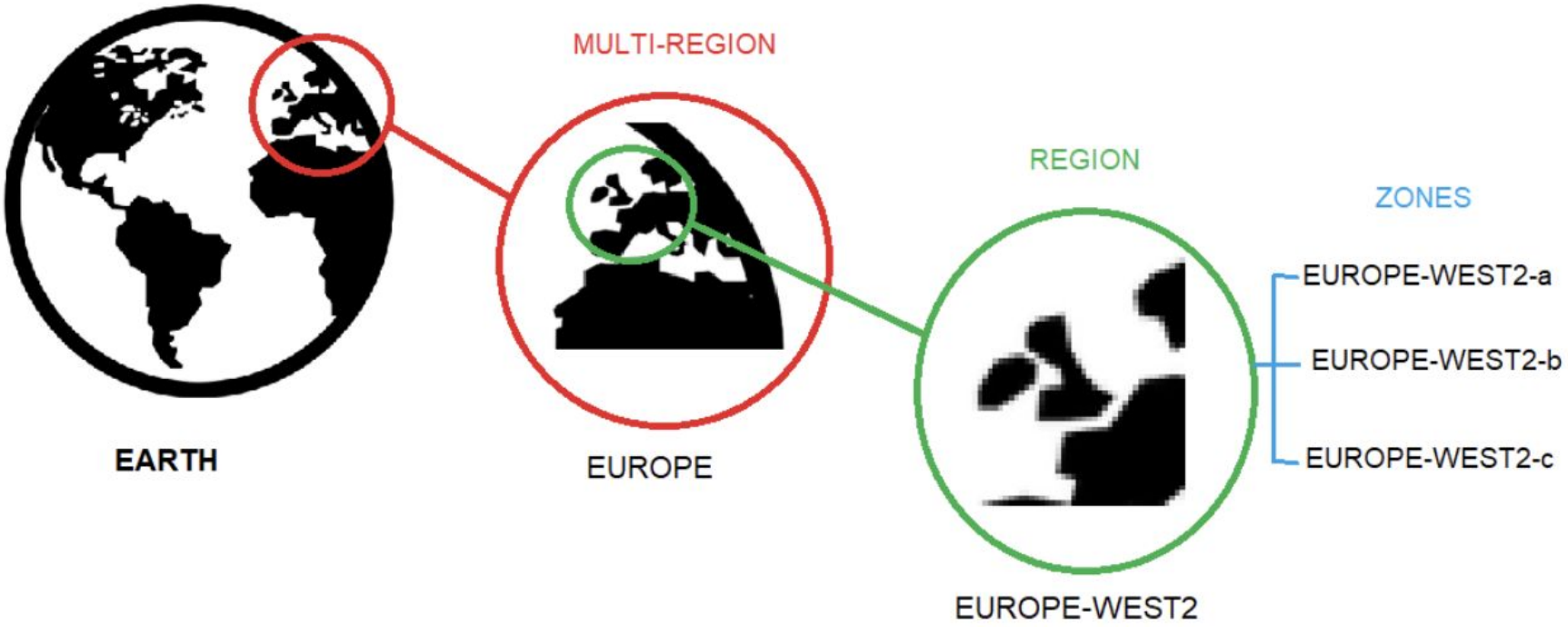
OS
Environment
Subnetwork
Boot Disk

Data Disk

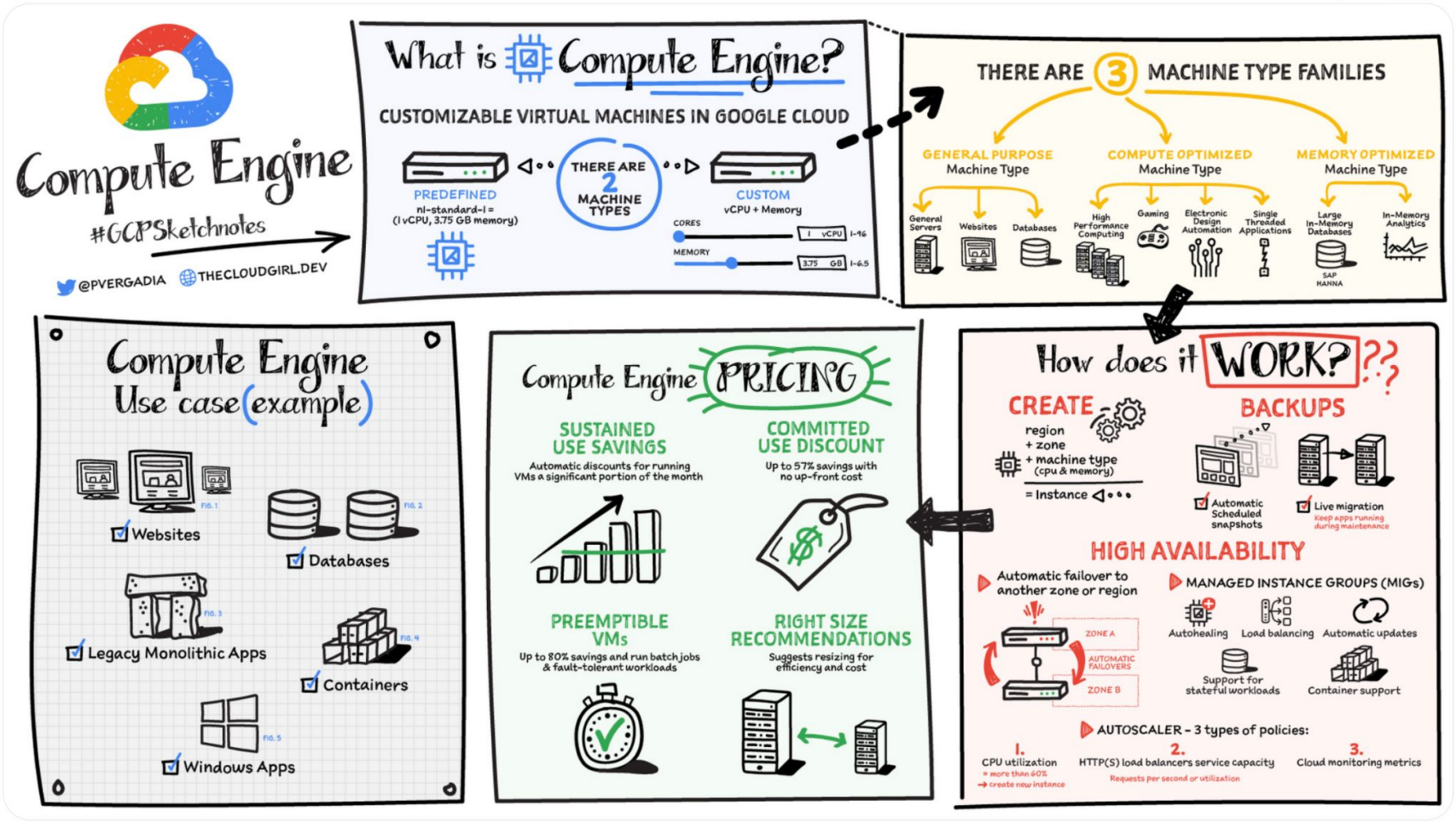
Permissions

Source: XXX

Regions / Zones



Source: https://www.geeksforgeeks.org/what-is-google-cloud-platform-gcp/

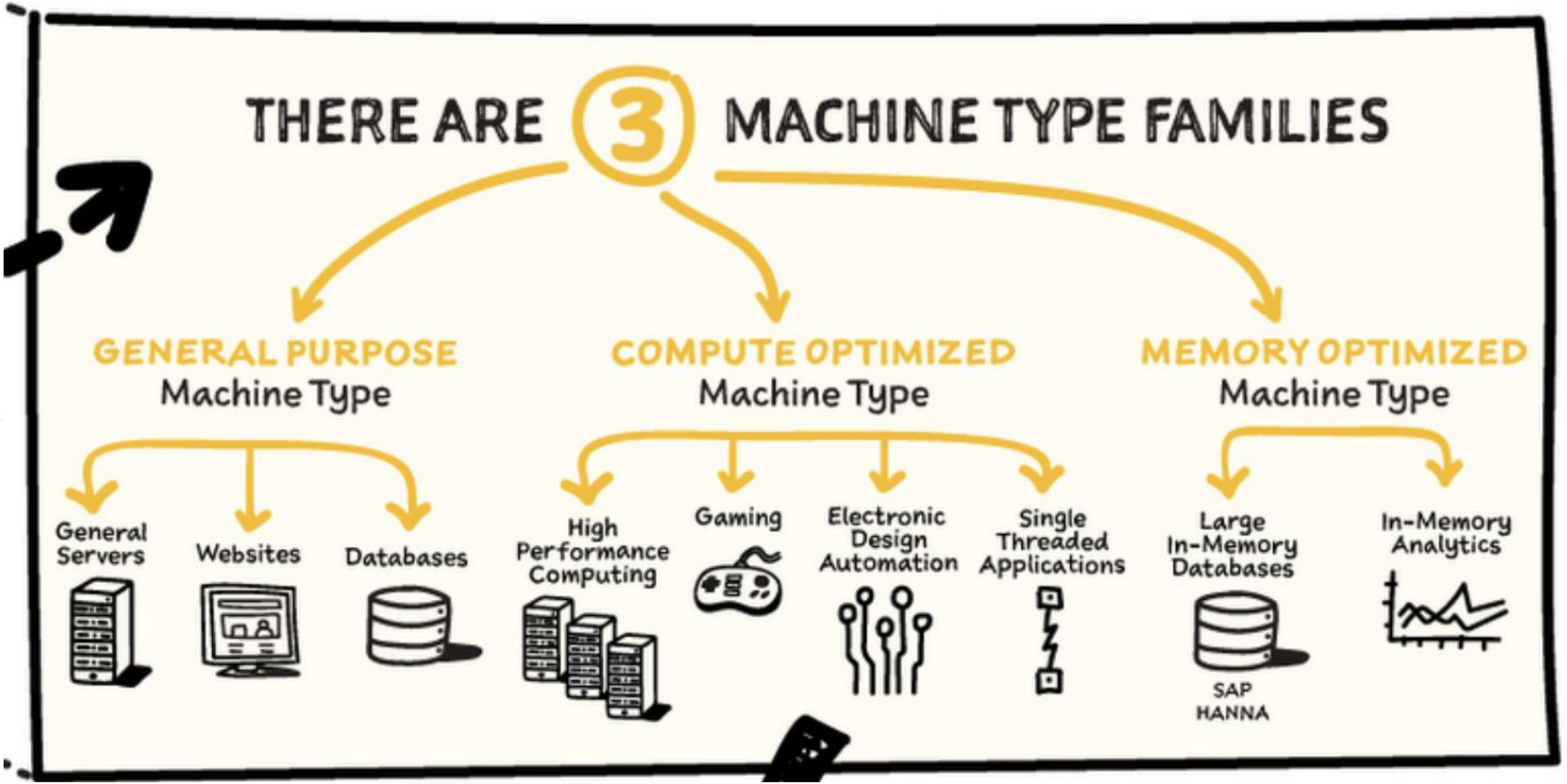


Powered by simplilearn

Source:
https://cloud.google.com/blog/topics/developers-practitioners/what-compute-engine-use-cases-security-pricing-and-more

Caltech | Center for Technology & Management Education

Machine Families

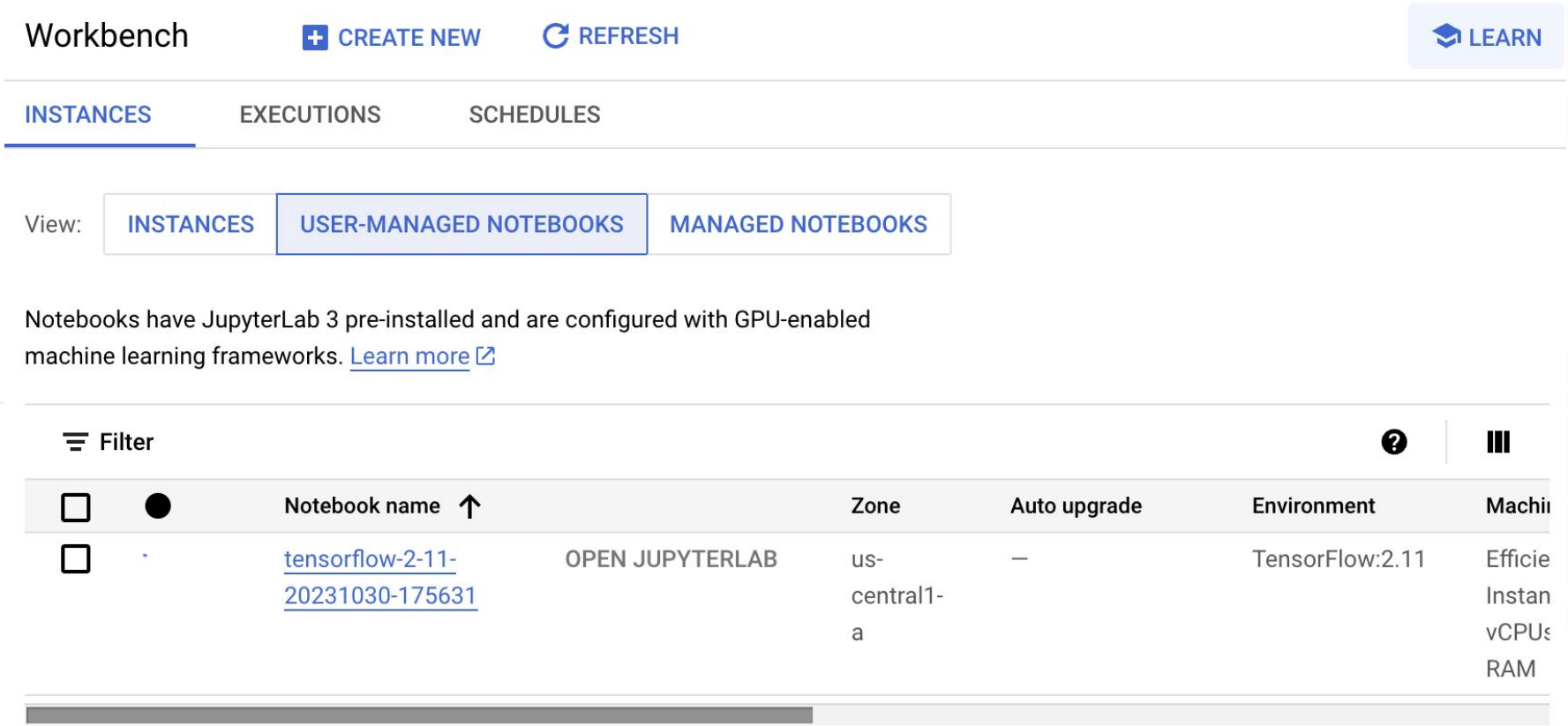


Source:
https://cloud.google.com/blog/topics/developers-practitioners/what-compute-engine-use-cases-security-pricing-and-more

Machine Types

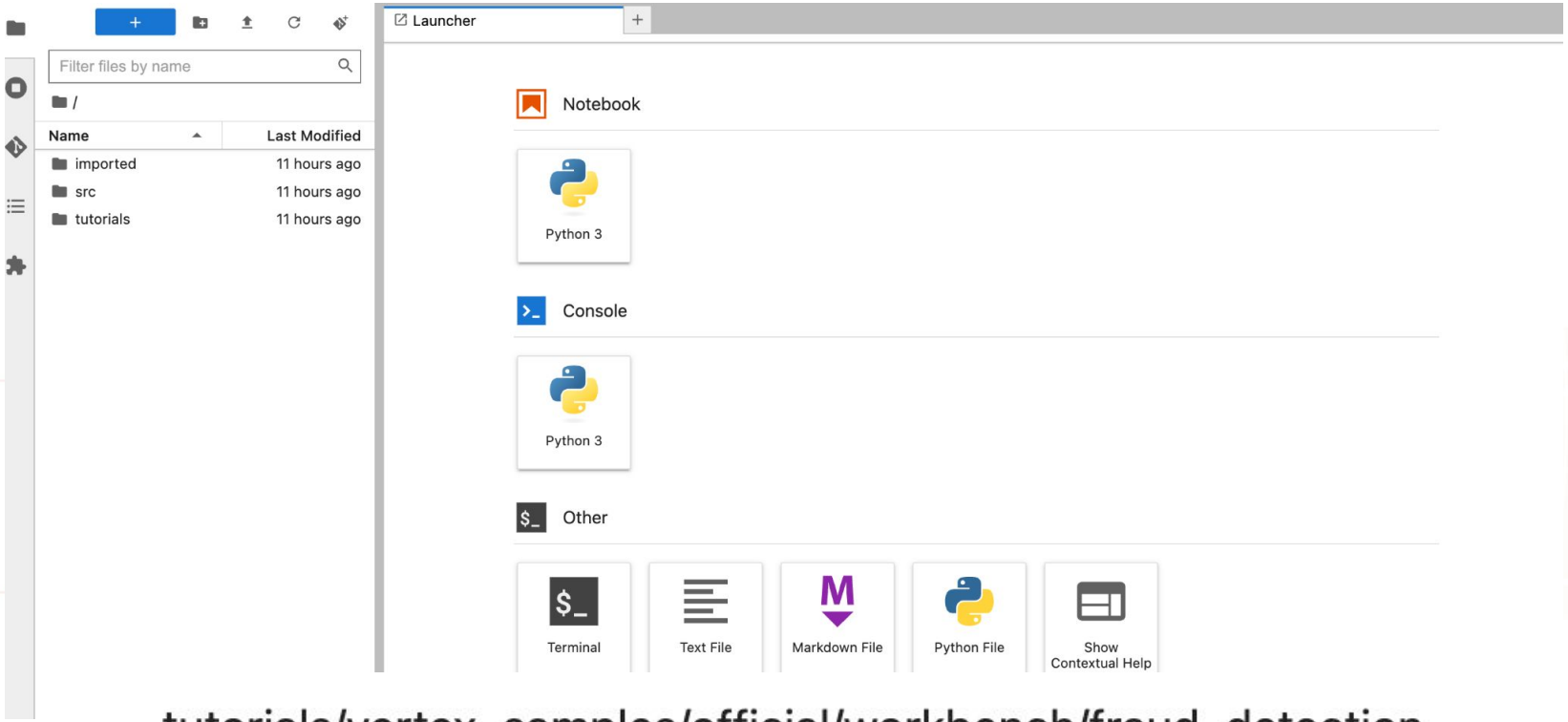
Workload type						
General-purpose workloads				Compute-optimized	Memory-optimized	Accelerator-optimized
E2	N2, N2D, N1	C3, C3D	Tau T2D, Tau T2A	H3, C2, C2D	M3, M2, M1	A2, G2
Day-to-day computing at a lower cost	Balanced price/performance across a wide range of machine types	Consistently high performance for a variety of workloads	Best per-core performance/cost for scale-out workloads	Ultra high performance for compute-intensive workloads	Highest memory to compute ratios for memory-intensive workloads	Optimized for accelerated high performance computing workloads

Source: https://cloud.google.com/compute/docs/machine-resource


Workbench
CREATE NEW
REFRESH
LEARN
INSTANCES
EXECUTIONS
SCHEDULES
View:
INSTANCES
USER-MANAGED NOTEBOOKS
MANAGED NOTEBOOKS
Notebooks have JupyterLab 3 pre-installed and are configured with GPU-enabled machine learning frameworks. Learn more
Filter
Notebook name
Zone
Auto upgrade
Environment
Machi
tensorflow-2-11-20231030-175631
OPEN JUPYTERLAB
us-central1-a
—
TensorFlow:2.11
Efficie
Instan
vCPUs
RAM

Wait for instance to spin up (Water Break)

Jupyterlab



tutorials/vertex_samples/official/workbench/fraud_detection

Installation

Install the following packages required to execute this notebook.

```
! pip3 install --upgrade --quiet google-cloud-aiplatform \
                                witwidget \
                                fsspec \
                                gcsfs
! pip3 install --quiet scikit-learn==1.2 \
                       protobuf==3.20.1
```

Before you begin

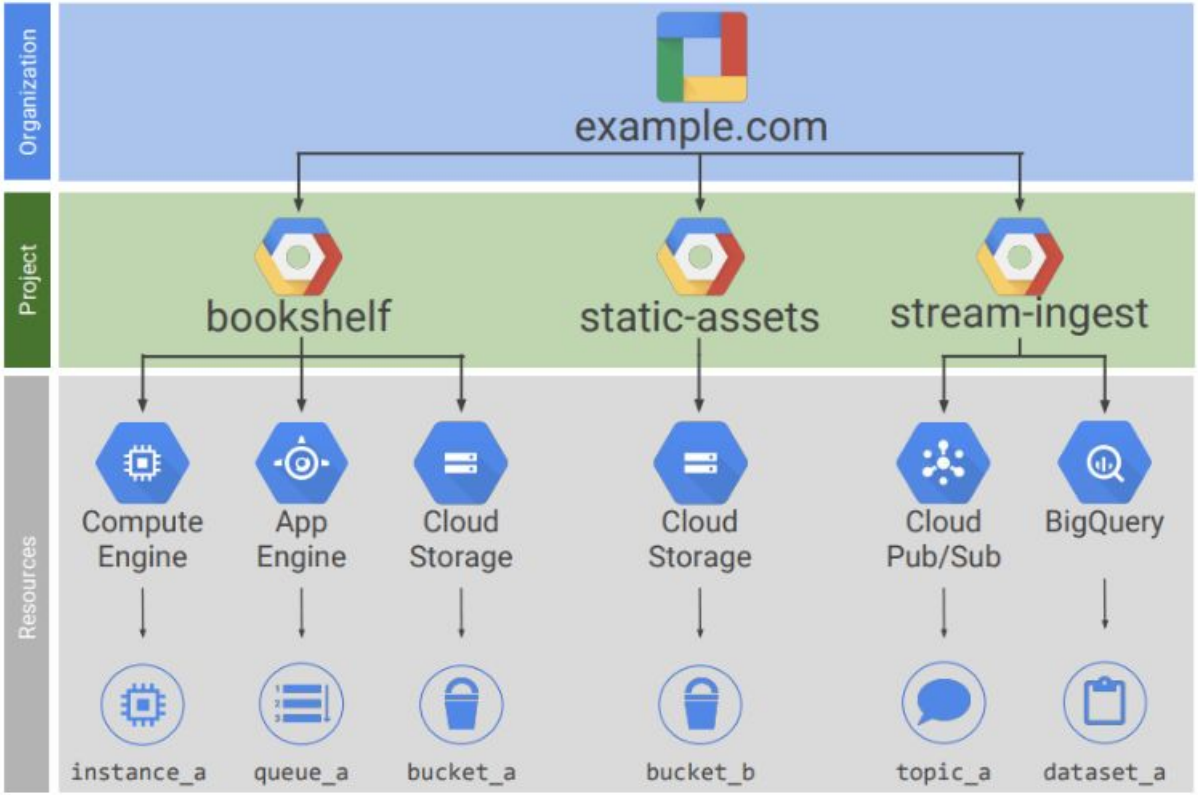
Set up your Google Cloud project

The following steps are required, regardless of your notebook environment.

1. Select or create a Google Cloud project. When you first create an account, you get a $300 free credit towards your compute/storage costs.
2. Make sure that billing is enabled for your project.
3. Enable the Vertex AI API.
4. If you are running this notebook locally, you need to install the Cloud SDK.

How to create a Project

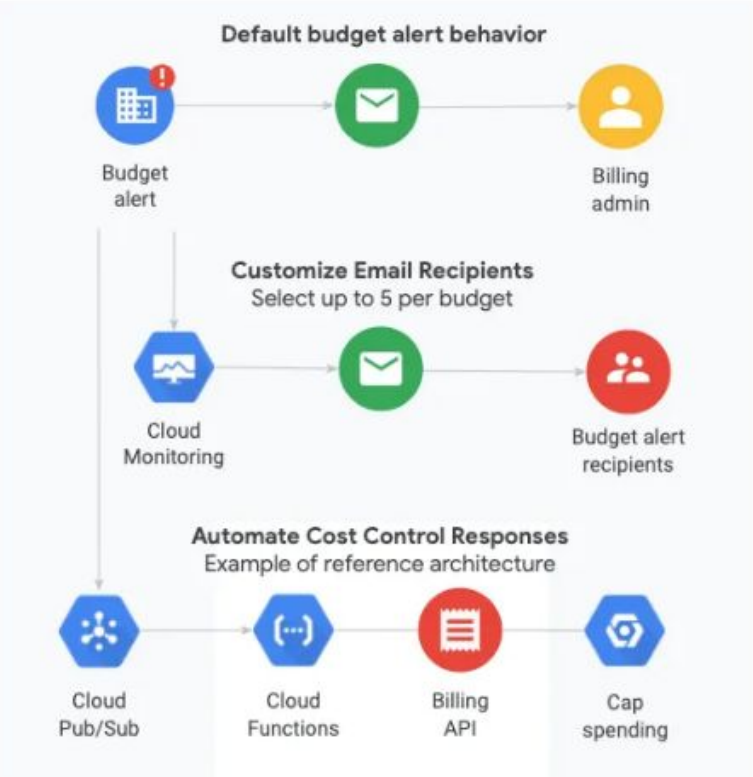
Creating a Google Project



Source: https://k21academy.com/google-cloud/google-cloud-projects/

Enable Billing

Create billing alerts [tutorial]

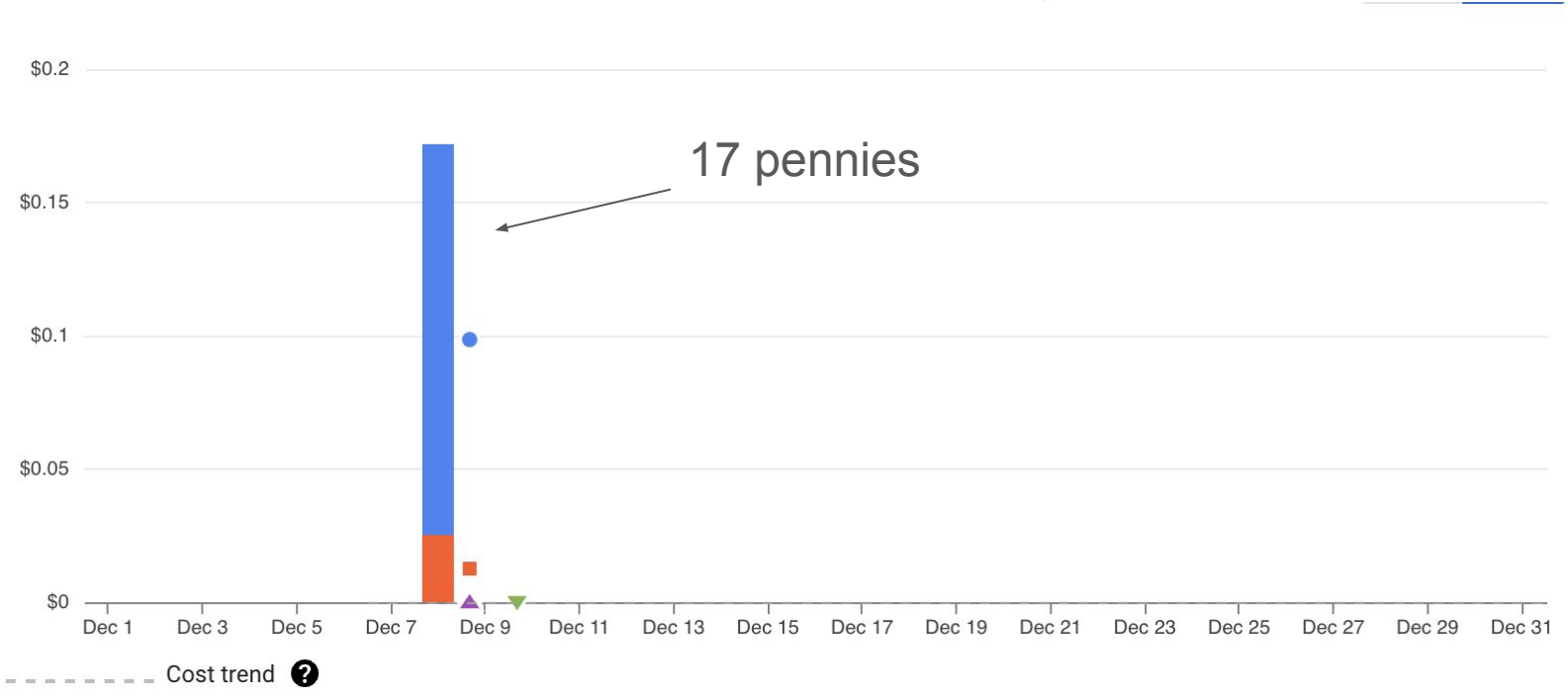


Google Cloud Budget Alerting Options

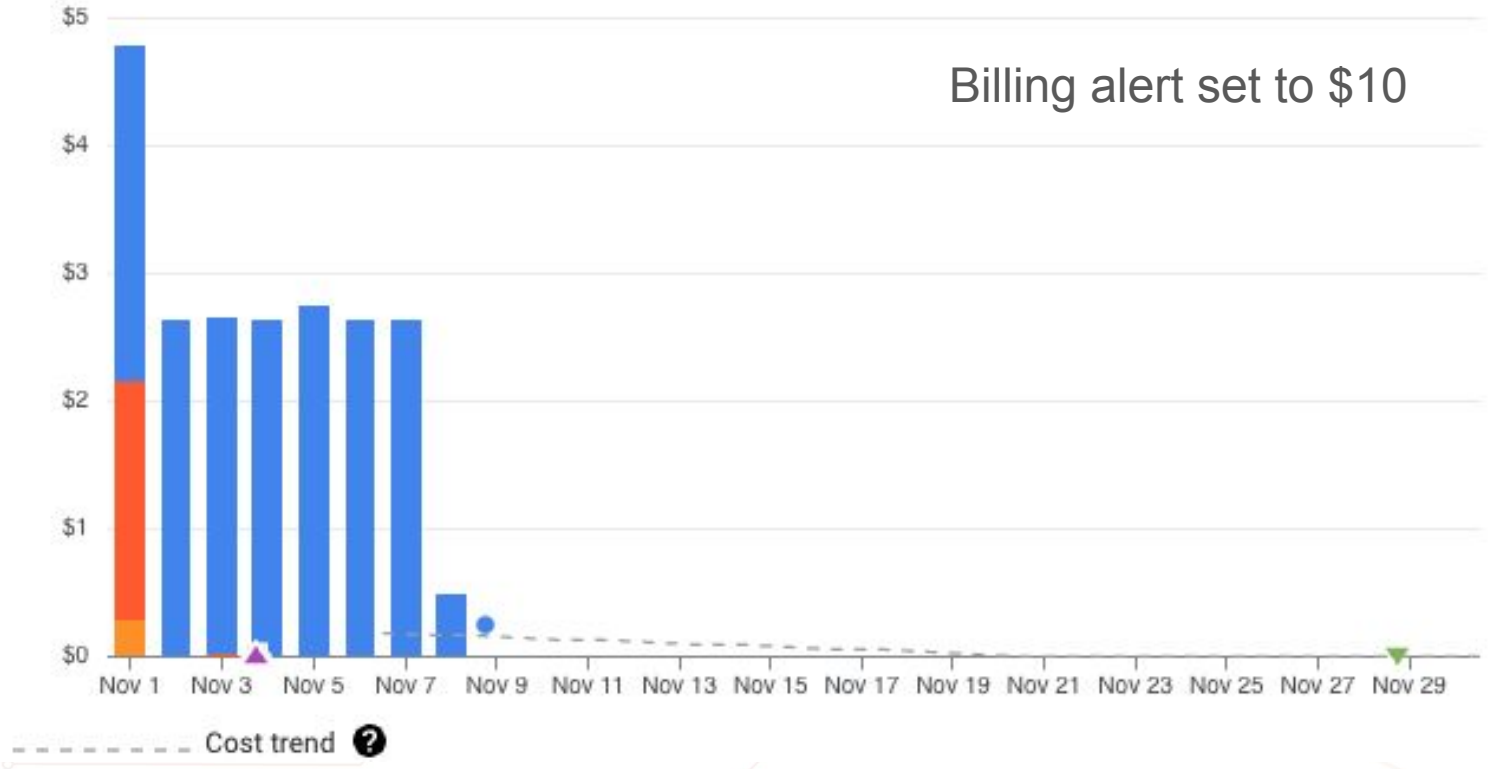
I do mine @ ~$10

Source: https://medium.com/google-cloud/google-cloud-billing-budgets-and-alerts-fc707ac1c2e4

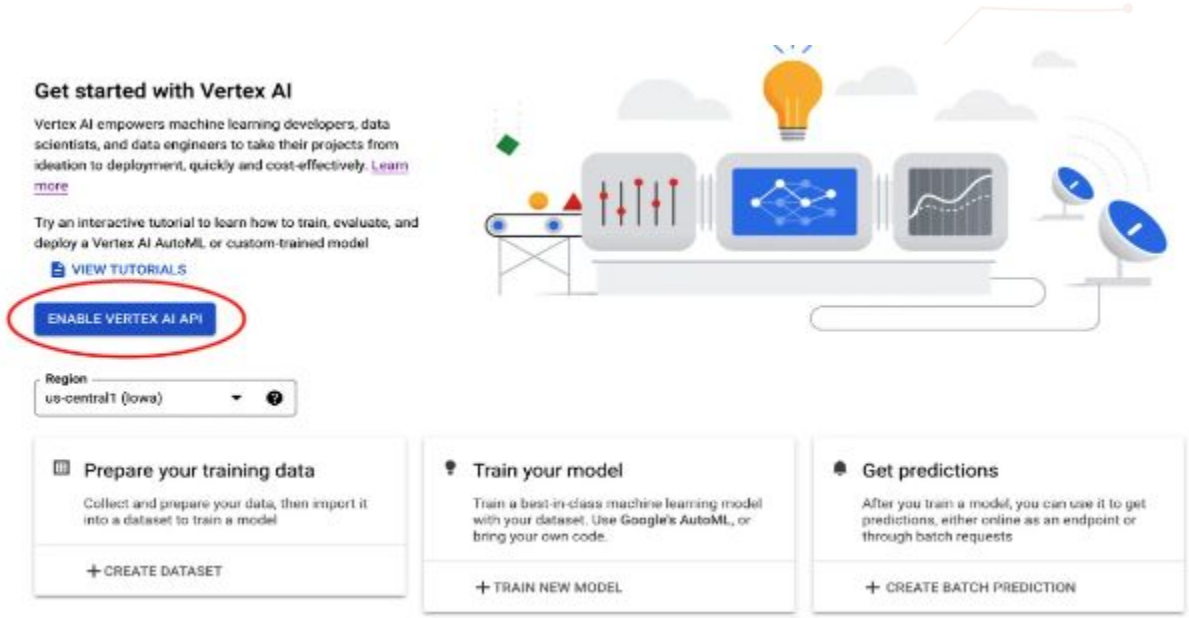
Cost of running this lab



Cost if you don't turn things off



Enable APIs



Source: https://subscription.packtpub.com/book/data/9781803233727/app3/app3lvl1sec02/vertex-ai-enabling-its-api

Back to Vertex AI

Setting the Project ID

Set your project ID

If you don't know your project ID, try the following:

- Run `gcloud config list`
- Run `gcloud projects list`
- See the support page: Locate the project ID

```
PROJECT_ID = "[your-project-id]"  # @param {type:"string"}

# set the project id
! gcloud config set project $PROJECT_ID
```

Project ID = caltech-class

Create buckets

Create a Cloud Storage bucket

Create a storage bucket to store intermediate artifacts such as datasets.

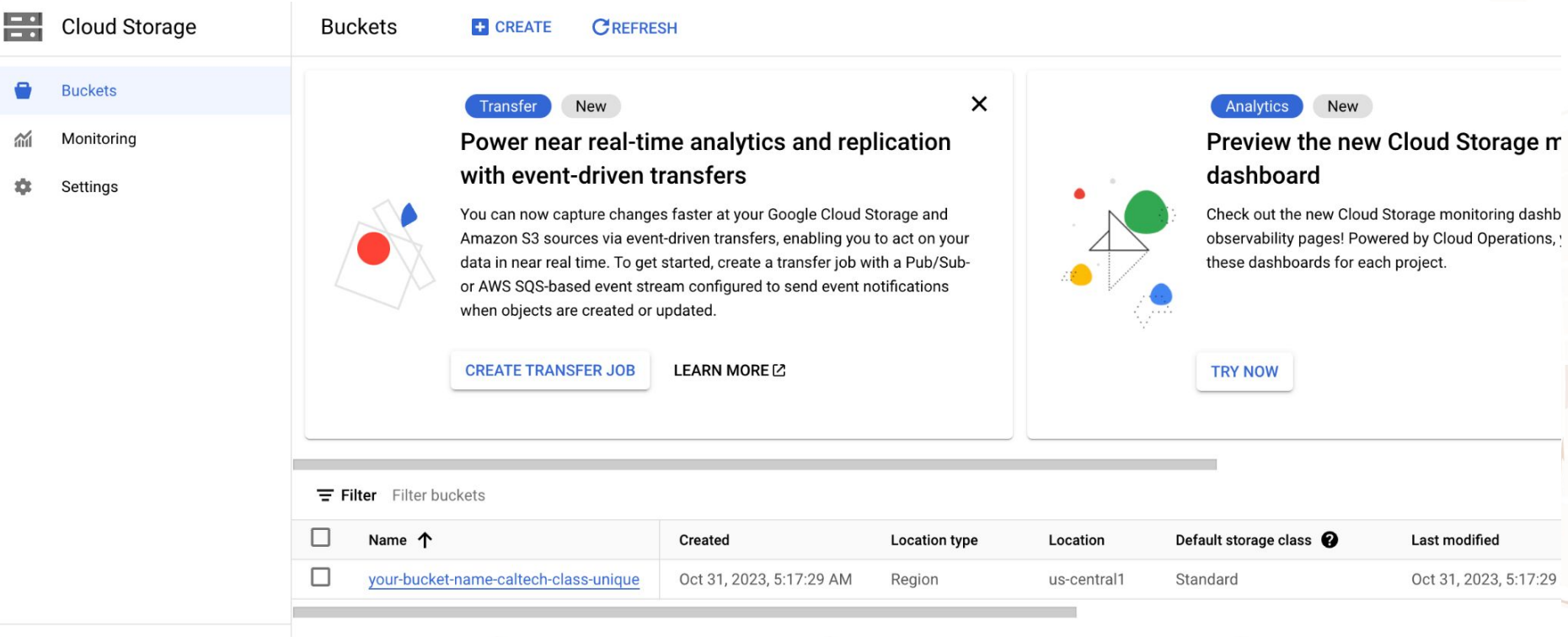
[6]:
```
BUCKET_URI = f"gs://your-bucket-name-{PROJECT_ID}-unique"  # @param {type:"string"}
```

Only if your bucket doesn't already exist: Run the following cell to create your Cloud Storage bucket.

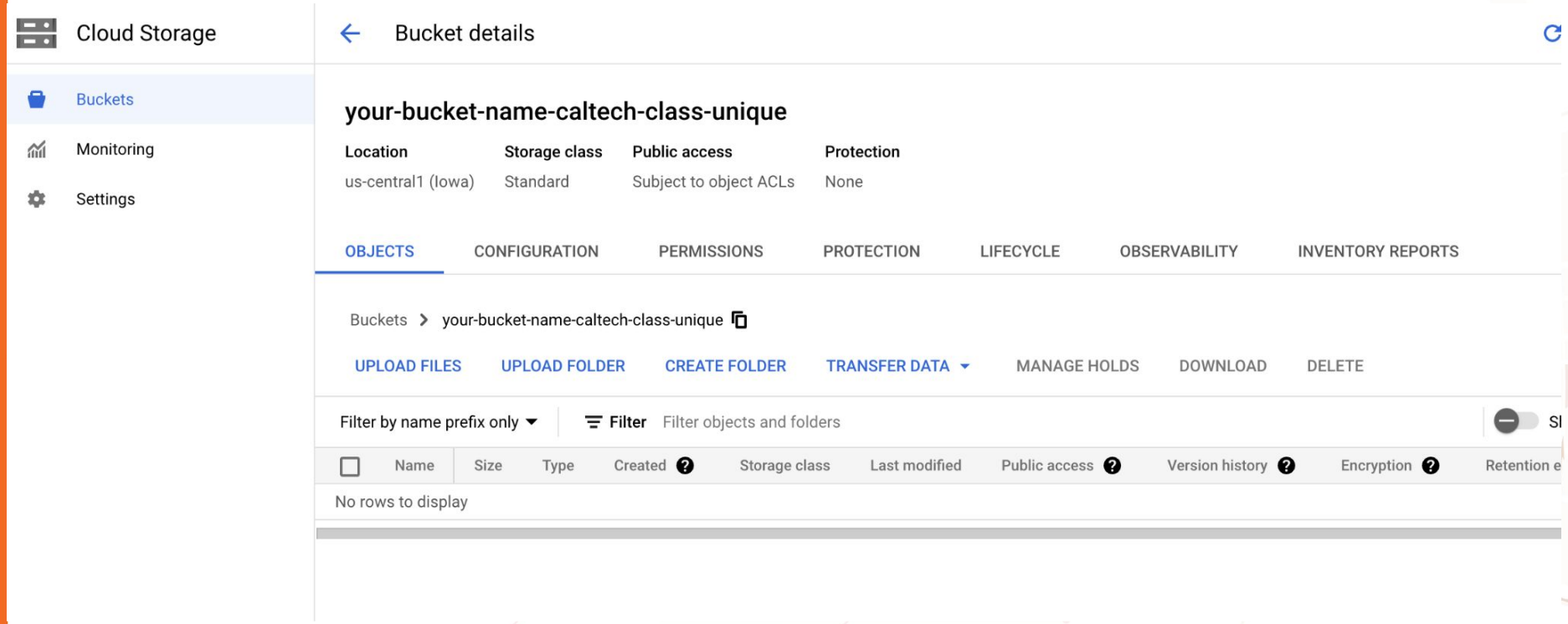
[7]:
```
! gsutil mb -l {REGION} -p {PROJECT_ID} {BUCKET_URI}
```

```
Creating gs://your-bucket-name-caltech-class-unique/...
```

Check on new bucket



Check on new bucket



Import libraries

[*]:
```
import os
import pickle
import sys
import warnings

import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from google.cloud import aiplatform, storage
from IPython.display import display
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import (average_precision_score, classification_report,
                             confusion_matrix, f1_score)
from sklearn.model_selection import train_test_split
from witwidget.notebook.visualization import WitConfigBuilder, WitWidget

warnings.filterwarnings("ignore")
```

Initialize Vertex AI SDK for Python

Initialize the Vertex AI SDK for Python for your project.

[9]:
```
aiplatform.init(project=PROJECT_ID, location=REGION, staging_bucket=BUCKET_URI)
```

Load dataset

Load the dataset from the public csv file path using Pandas.

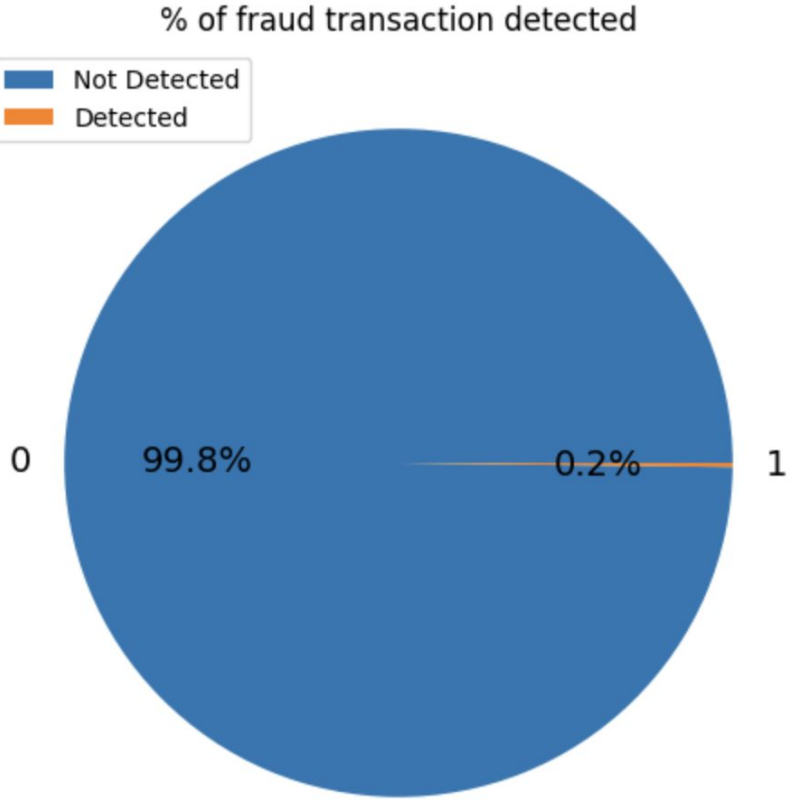
```
[ ]:
# set the dataset path
DATASET_SOURCE_PATH = "gs://cloud-samples-data/vertex-ai/managed_notebooks/fraud_detection/fraud_detection_data.csv"
# read the csv data using pandas
df = pd.read_csv(DATASET_SOURCE_PATH)
```

Analyze the dataset

Take a quick look at the dataset and the number of rows

```
# print the shape of dataframe
print("shape : ", df.shape)
# display the dataframe
df.head()
```

Imbalanced data



Data prep

Prepare data for modeling

To prepare the dataset for training, a few columns need to be dropped that contain unique data ('nameOrig','nam ('isFlaggedFraud'). The categorical field "type" which describes the type of transaction and is important for fraud hot encoded.

]:
```
# drop the unnecessary fields
df.drop(["nameOrig", "nameDest", "isFlaggedFraud"], axis=1, inplace=True)
# encode the "type" field
X = pd.concat([df.drop("type", axis=1), pd.get_dummies(df["type"])], axis=1)
X.head()
```

Remove the outcome variable from the training data.

]:
```
# copy the target data
y = X[["isFraud"]]
# remove the target field from the features
X = X.drop(["isFraud"], axis=1)
```

Train-test strategy

Split the data and assign 70% for training and 30% for testing.

For splitting, you specify the following parameters to Sklearn's `train_test_split` method:

- `*arrays` : The feature array(X) and the target array(y).
- `test_size` : Percentage(float) or number(integer) of test samples.
- `random_state` : Controls the shuffling applied to the data before applying the split. Pass an int for reproducible output across multiple function calls.
- `stratify` : If none, no stratified sampling is performed.

As the data is imbalanced, you use stratified sampling while splitting. Learn more about stratified sampling and other parameters for train-test-splitting.

Model selection

Fit a Random Forest model

Fit a simple Random Forest classifier on the preprocessed training dataset.

Note: Setting `n_jobs` to -1 while defining the `RandomForestClassifier` object allows it to parallelize the training process using all processors.

Learn more about Random Forest algorithm and Sklearn's RandomForestClassifier.

```
# create a randomforestclassifier object
forest = RandomForestClassifier(n_jobs=-1, verbose=1)
# fit the model on the data
forest.fit(X_train, y_train)
```

This will take about 7 min

Model training

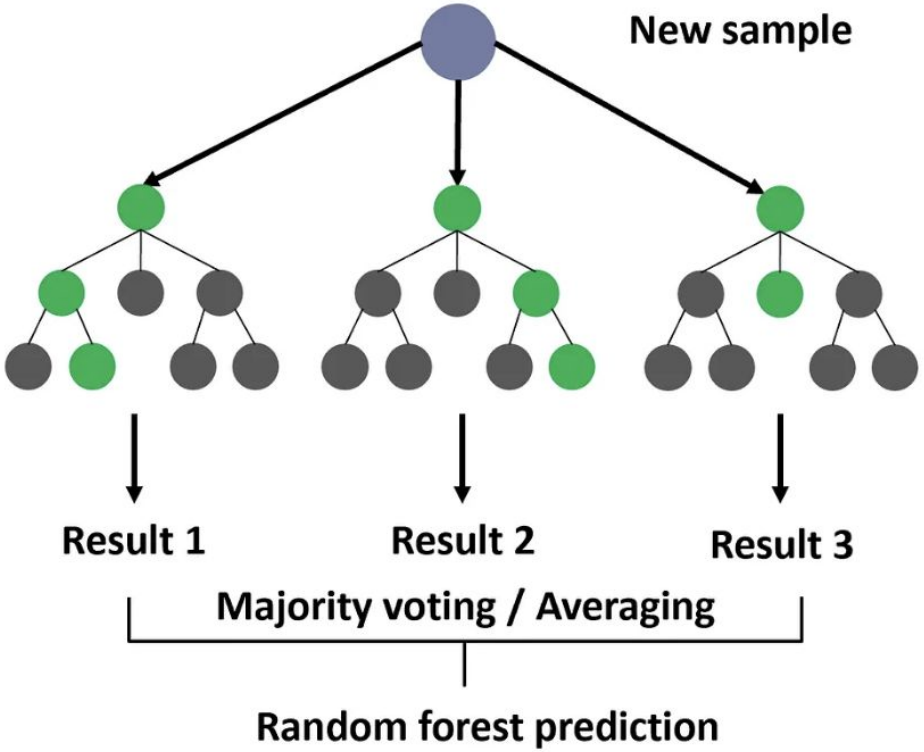
```
# create a randomforestclassifier object
forest = RandomForestClassifier(n_jobs=-1, verbose=1)
# fit the model on the data
forest.fit(X_train, y_train)
```

```
[Parallel(n_jobs=-1)]: Using backend ThreadingBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done  42 tasks      | elapsed:  3.2min
[Parallel(n_jobs=-1)]: Done 100 out of 100 | elapsed:  7.2min finished
```

▾ RandomForestClassifier

```
RandomForestClassifier(n_jobs=-1, verbose=1)
```

Random Forest Overview (while training)



Source: https://medium.com/@roiyeho/random-forests-98892261dc49

Analyze results

```
AP : 0.9376763239087629
F1 - score : 0.8665307047360074
Confusion_matrix :
[[1906285      37]
 [    552    1912]]
classification_report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1906322
           1       0.98      0.78      0.87      2464

    accuracy                           1.00   1908786
   macro avg       0.99      0.89      0.93   1908786
weighted avg       1.00      1.00      1.00   1908786
```

What do we use for classification?

Accuracy?

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100$$

Going beyond accuracy

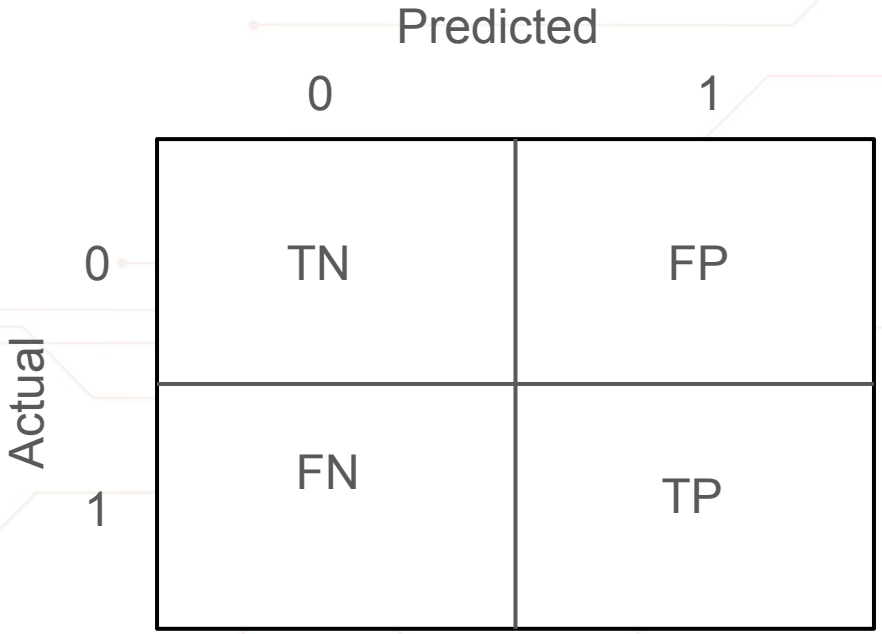
Classification Evaluation Metrics

Metric	Formula	Evaluation focus
Accuracy	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$	Overall effectiveness of a classifier
Precision	$PRC = \frac{TP}{TP + FP}$	Class agreement of the data labels with the positive labels given by the classifier
Sensitivity	$SNS = \frac{TP}{TP + FN}$	Effectiveness of a classifier to identify positive labels. Also called true positive rate (TPR)
Specificity	$SPC = \frac{TN}{TN + FP}$	How effectively a classifier identifies negative labels. Also called true negative rate (TNR)
F_1 score	$F_1 = 2\frac{PRC \cdot SNS}{PRC + SNS}$	Combination of precision (PRC) and sensitivity (SNS) in a single metric
Geometric mean	$GM = \sqrt{SNS \cdot SPC}$	Combination of sensitivity (SNS) and specificity (SPC) in a single metric
Area under (ROC) curve	$AUC = \int_0^1 SNS \cdot dSPC$	Combined metric based on the receiver operating characteristic (ROC) space (*Powers, 2011*)

Source: https://www.researchgate.net/figure/Classification-performance-metrics-based-on-the-confusion-matrix_tbl3_324952663

Simple, right?

The **Confusion Matrix**: Where it all begins



Great 35 minute video by Data School

Example: Deadly Disease classification model

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Overall, how often is the classifier correct?

(TP+TN)/total = (100+50)/165 = 0.91

91% accurate

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	TN = 50	FP = 10
Actual: YES	FN = 5	TP = 100


Let's celebrate !!
getclippy.co

Not so fast buddy

- What if 91% accuracy is a terrible model?
- What if we are inaccurately classifying a large group of people?
- How much does misclassification hurt us/the organization/business/project?
- Do we care more about misclassifying False Positives?
- What about False Negatives?

The Accuracy Paradox

Example [edit]

For example, a city of 1 million people has ten terrorists. A profiling system results in the following confusion matrix:

Predicted class / Actual class	Fail	Pass	Sum
Fail	**10**	0	10
Pass	**990**	999000	999990
Sum	1000	999000	1000000

Even though the accuracy is $\frac{10 + 999000}{1000000} \approx 99.9\%$, 990 out of the 1000 positive predictions are incorrect. The precision of $\frac{10}{10 + 990} = 1\%$ reveals its poor performance. As the classes are so unbalanced, a better metric is the F1 score $= \frac{2 \times 0.01 \times 1}{0.01 + 1} \approx 2\%$ (the recall being $\frac{10 + 0}{10} = 1$).

Wait… what?

Precision and Recall

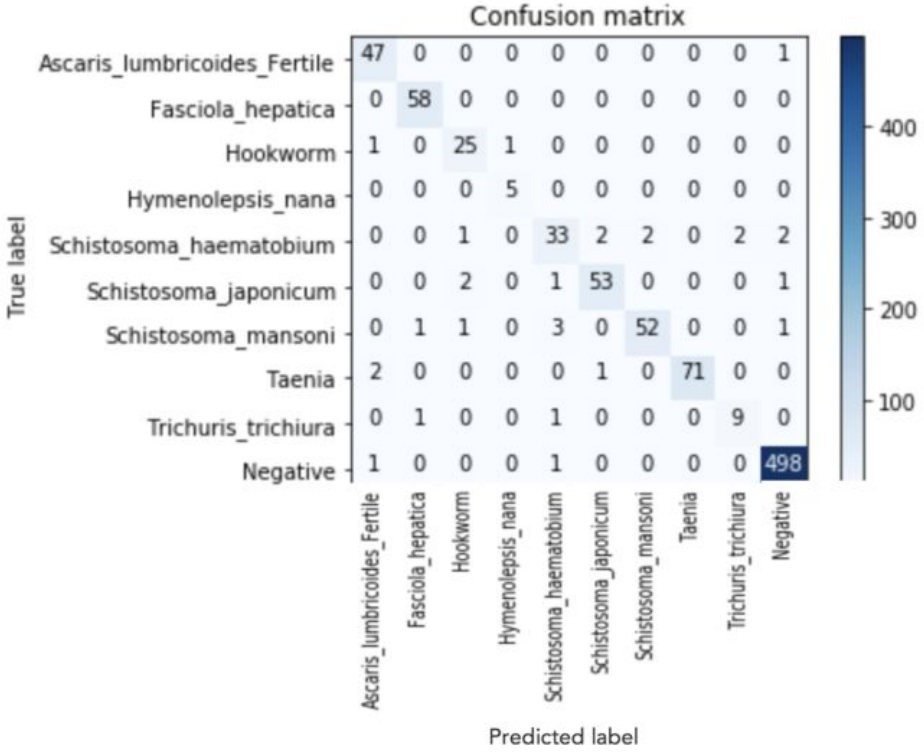
- **Recall:** When it's actually yes, how often does it predict yes?
  - TP/actual yes = 100/105 = 0.95
- **Precision:** When it predicts yes, how often is it correct?
  - TP/predicted yes = 100/110 = 0.91

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

n=165	Predicted: NO	Predicted: YES
Actual: NO	TN = 50	FP = 10
Actual: YES	FN = 5	TP = 100

Multiclass Confusion Matrix



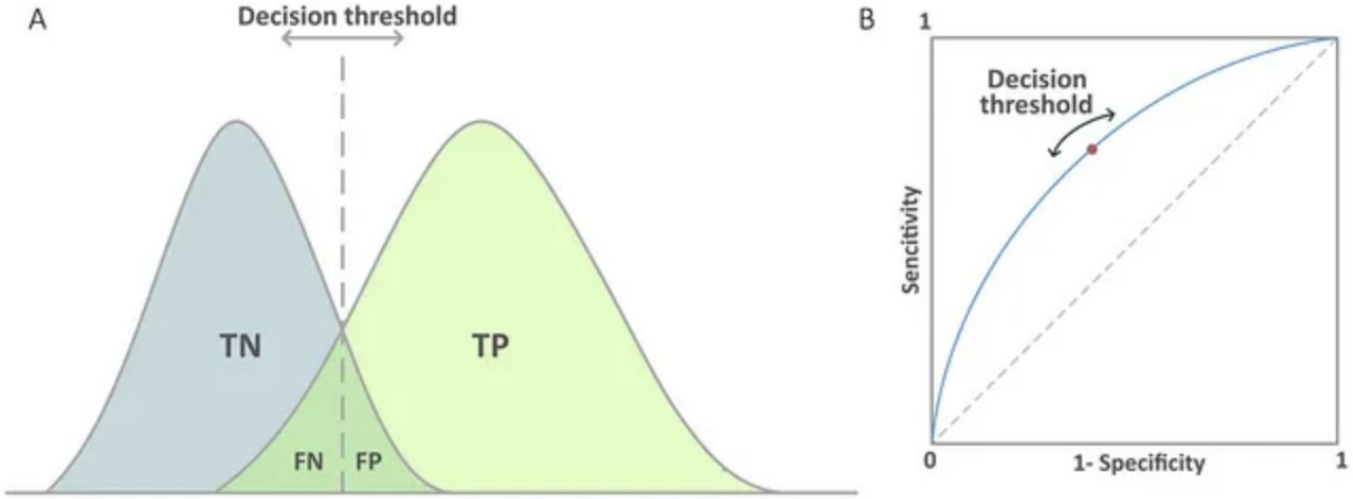
Holistically, what are we talking about?

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Precision} = \frac{TP}{TP+FP}$$

What about some harmony?

$$F1\,\text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Probability thresholds

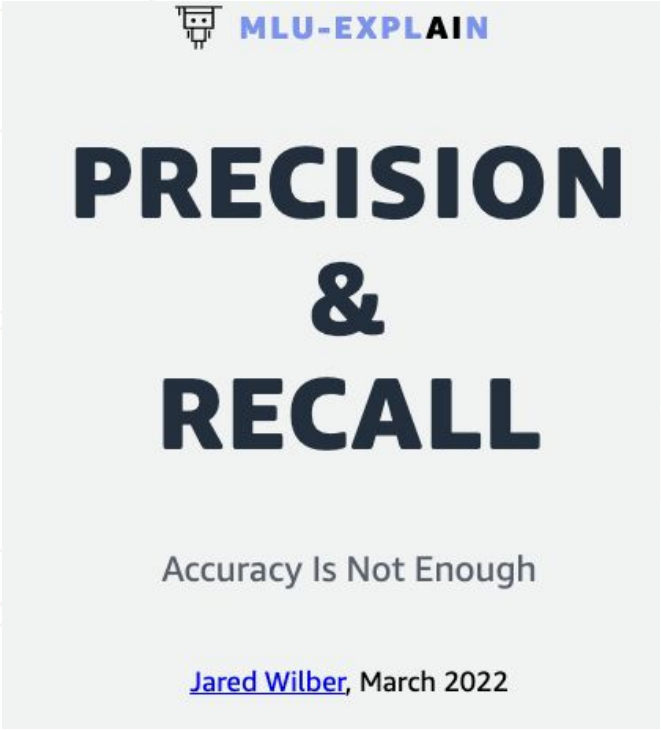


The goal of this post is to outline how to move the decision threshold to the left in Figure A, reducing false negatives and maximizing sensitivity.

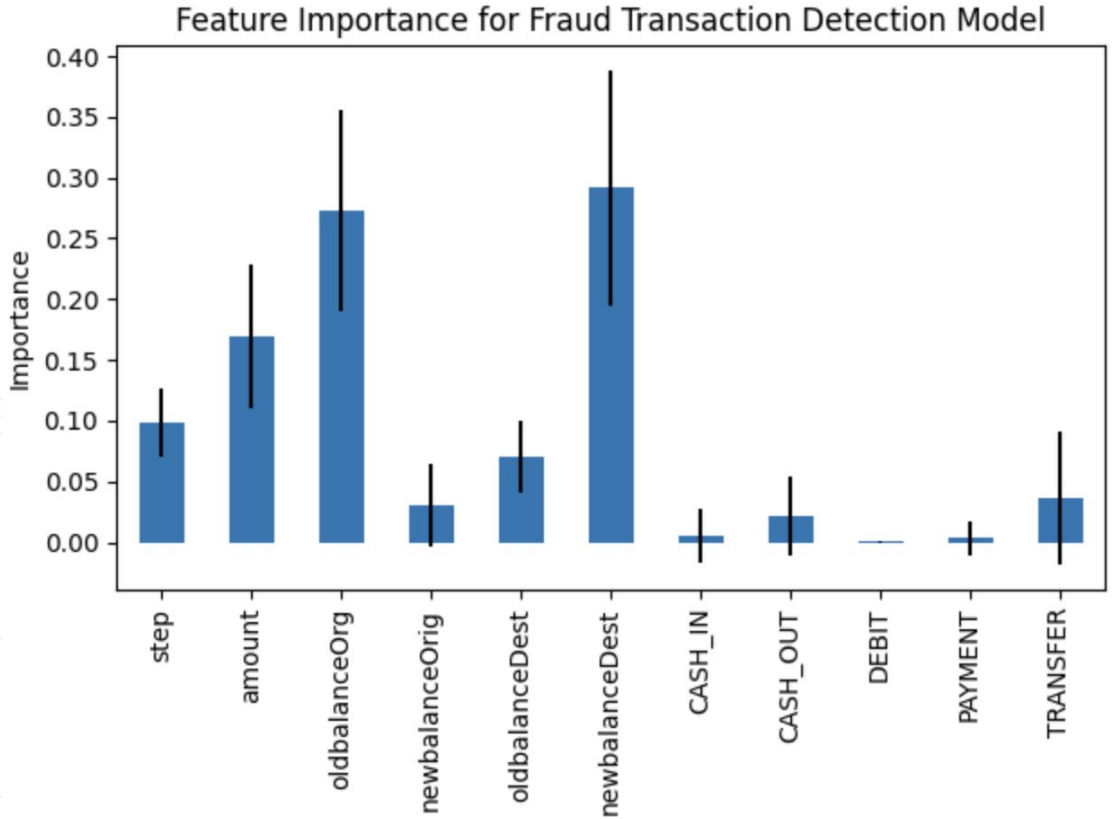
Source: https://towardsdatascience.com/fine-tuning-a-classifier-in-scikit-learn-66e048c21e65

Let’s check a cool site out

Dive in!



Analyze results

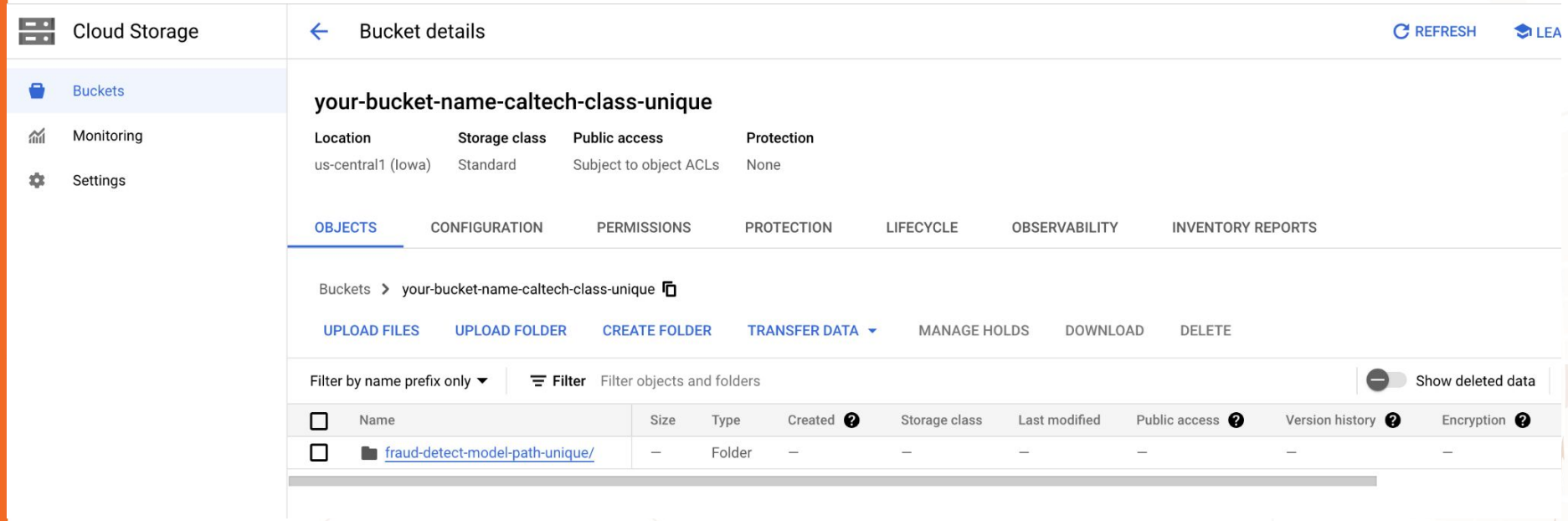


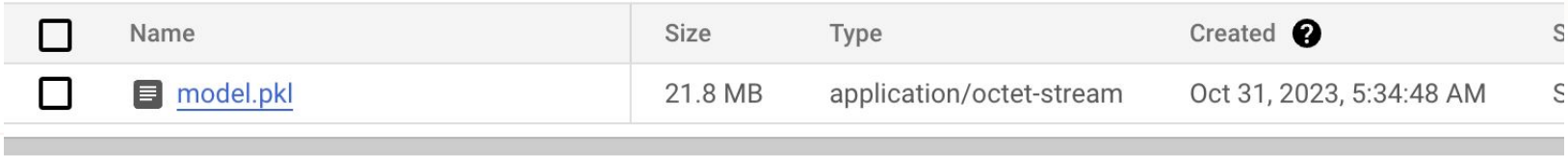
Save the model to Cloud Storage

Save your model to a pickle file and then, upload your model to Cloud Storage bucket. The uploaded model path is later used for creating a model in the Vertex AI Model Registry.

Note: You can also upload the model to Vertex AI Model Registry from your local environment using the latest Vertex AI SDK for Python.

Check your bucket





Create a model in Vertex AI

Set the parameters required for model creation in Vertex AI Model Registry.

```
# set model display name
MODEL_DISPLAY_NAME = "fraud-detection-model-unique"  # @param {type:"string"}
# set the GCS path to the model artifact
ARTIFACT_GCS_PATH = f"{BUCKET_URI}/{BLOB_PATH}"
# set the prediction container uri
SERVING_CONTAINER_IMAGE_URI = (
    "us-docker.pkg.dev/vertex-ai/prediction/sklearn-cpu.1-2:latest"
)
```

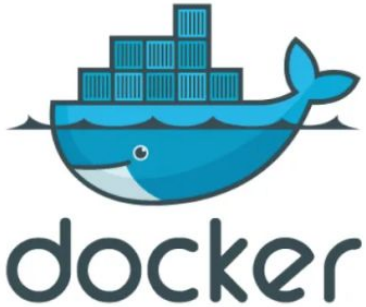
What the heck is a container?

What is a container?

→ Stores your application's code with files/libraries needed to run it on any machine

But why?

- More portable
- Solves the "works on my machine" situation
- Scales very easily when your app gets popular
- Fault tolerant → isolated from other infrastructure



"What Docker really does is separate the application code from infrastructure requirements and needs. It does this by running each application in an isolated environment called a 'container.' "

Source: Chinmay Shah

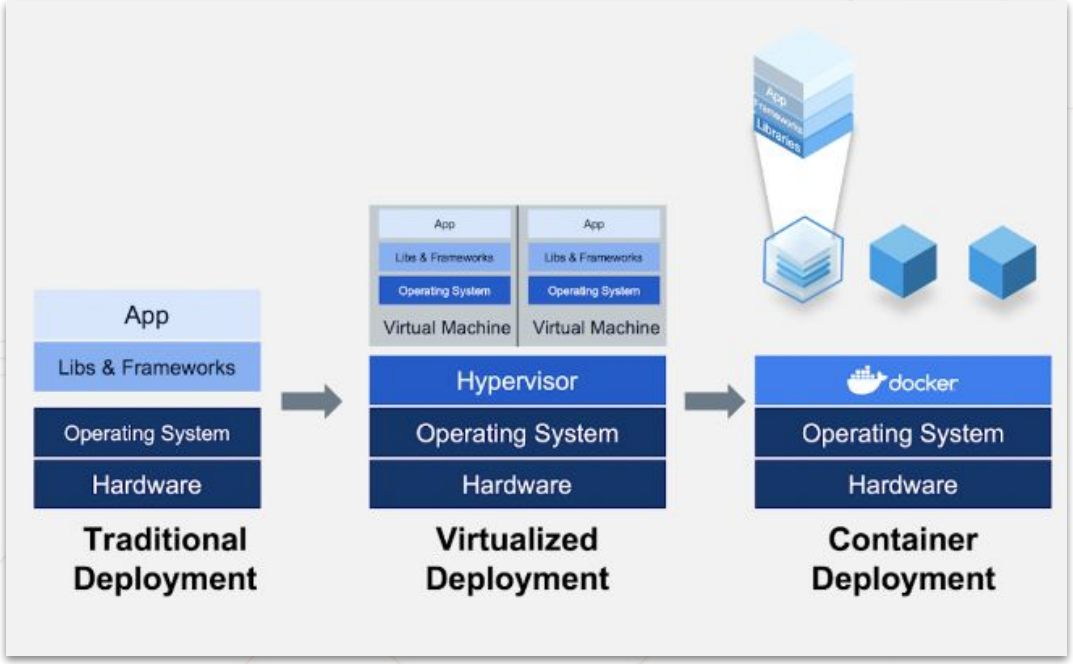
“It works on my computer”

The notebook won’t run on his colleagues computer!

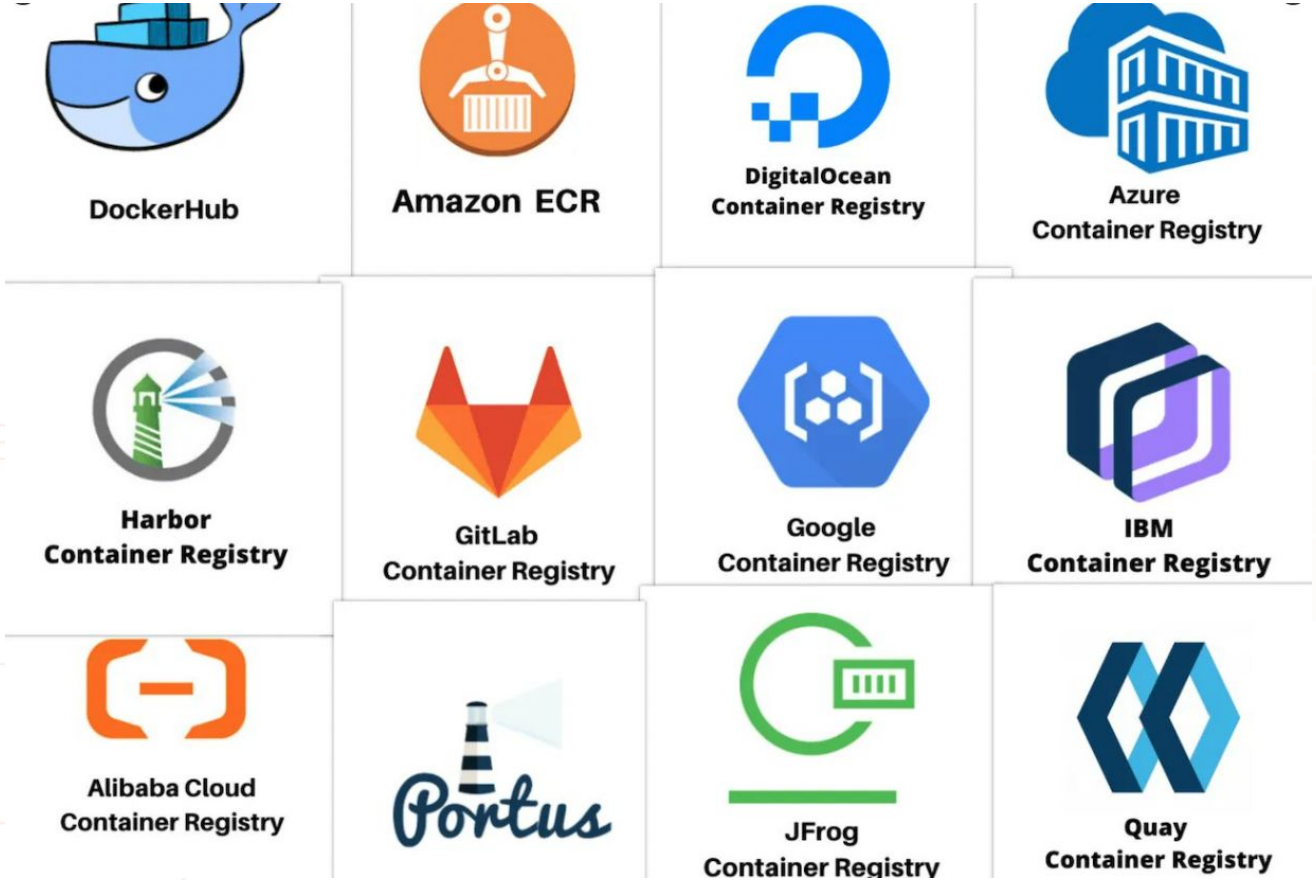
```
In [2]: import sys
        !{sys.executable} -m pip install --user numpy

de0e77425/numpy/distutils/command/build_src.py", line 378, in generate_sources
      source = func(extension, build_dir)
    File "/private/var/folders/5z/6mpx11111jx9spjrrht8k9y80000gn/T/pip-install-dduq83ml/numpy_8396f5e1bbf1412cae468cb
de0e77425/numpy/core/setup.py", line 456, in generate_config_h
      moredefs, ignored = cocache.check_types(config_cmd, ext, build_dir)
    File "/private/var/folders/5z/6mpx11111jx9spjrrht8k9y80000gn/T/pip-install-dduq83ml/numpy_8396f5e1bbf1412cae468cb
de0e77425/numpy/core/setup.py", line 50, in check_types
      out = check_types(*a, **kw)
    File "/private/var/folders/5z/6mpx11111jx9spjrrht8k9y80000gn/T/pip-install-dduq83ml/numpy_8396f5e1bbf1412cae468cb
de0e77425/numpy/core/setup.py", line 311, in check_types
      raise SystemError(
  SystemError: Cannot compile 'Python.h'. Perhaps you need to install python-dev|python-devel.
  ----------------------------------------
  ERROR: Failed building wheel for numpy
Failed to build numpy
ERROR: Could not build wheels for numpy, which is required to install pyproject.toml-based projects
WARNING: You are using pip version 21.3.1; however, version 22.0.3 is available.
You should consider upgrading via the '/Library/Frameworks/Python.framework/Versions/3.10/bin/python3 -m pip install
--upgrade pip' command.
```

Basic libraries are failing to install

Docker architecture

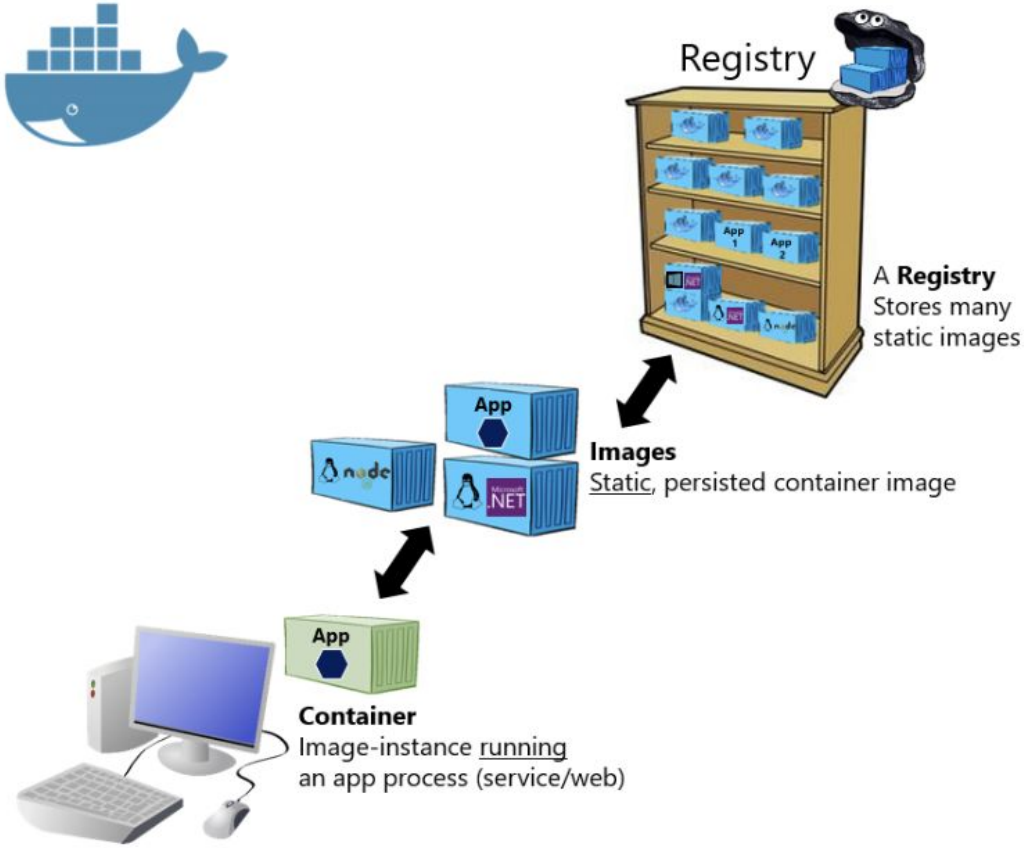


Lots of container registry options



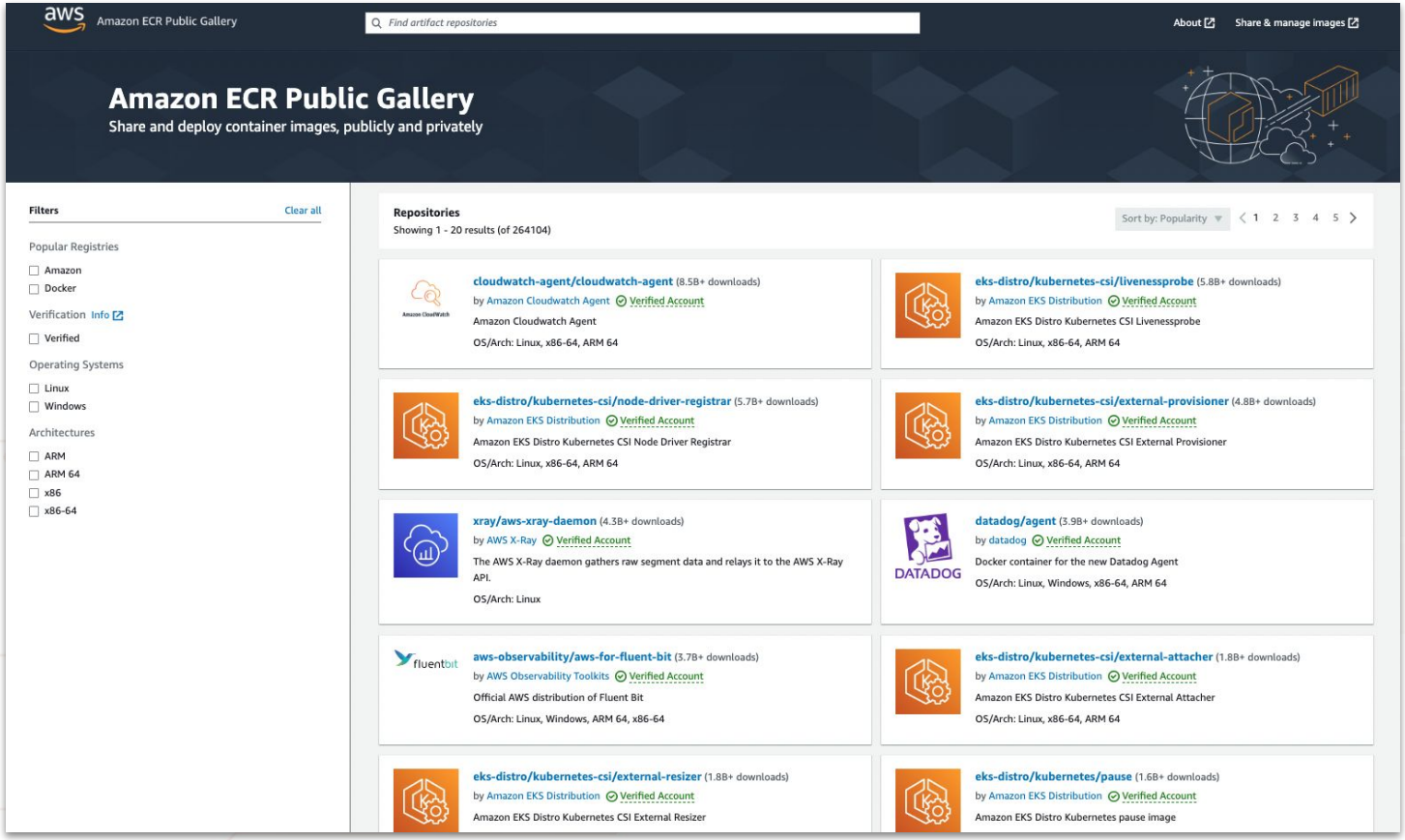
Source: https://twitter.com/razorops/status/1526782166524379139

Container Registry



Source: https://learn.microsoft.com/en-us/dotnet/architecture/microservices/container-docker-introduction/docker-containers-images-registries

Example container registry



What is Model Registry?

The Vertex AI Model Registry is a central repository where you can manage the lifecycle of your ML models. From the Model Registry, you have an overview of your models so you can better organize, track, and train new versions. When you have a model version you would like to deploy, you can assign it to an endpoint directly from the registry, or using aliases, deploy models to an endpoint.

Common workflow

There are many valid workflows for working in the Model Registry. To get started, you might want to follow the guidelines below to understand what you can do in the Model Registry and at what stage in your model-training journey.

- Import models to the Model Registry.
- Create new models, assign a model version the default alias, ready for production.
- Add other aliases, or labels to help you manage and organize your models and model versions.
- Deploy your models to an endpoint.
- Run batch prediction, and start your model evaluation pipeline.
- View your model details and view performance metrics from the model details page.

Source: GCP

Model Registry

Create a model resouce in Vertex AI using the `Model.upload` method.

Learn more about Vertex AI Model Registry.

```
# create a Vertex AI model resource
model = aiplatform.Model.upload(
    display_name=MODEL_DISPLAY_NAME,
    artifact_uri=ARTIFACT_GCS_PATH,
    serving_container_image_uri=SERVING_CONTAINER_IMAGE_URI,
)
# print the model's display name
print("Display name:\n", model.display_name)
# print the model's resource name
print("Resource name:\n", model.resource_name)
```

What is Model Registry?

MODEL DEVELOPMENT

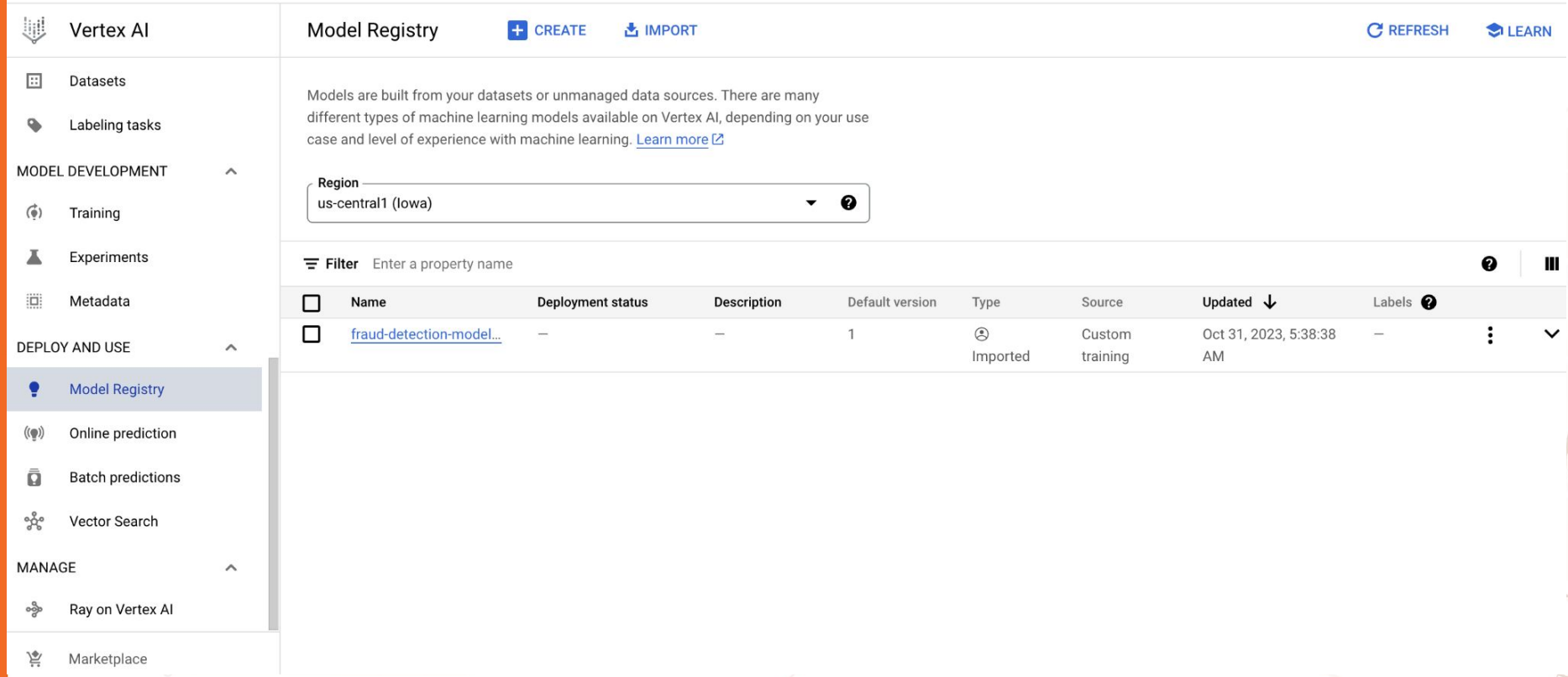
- Training
- Experiments
- Metadata

DEPLOY AND USE

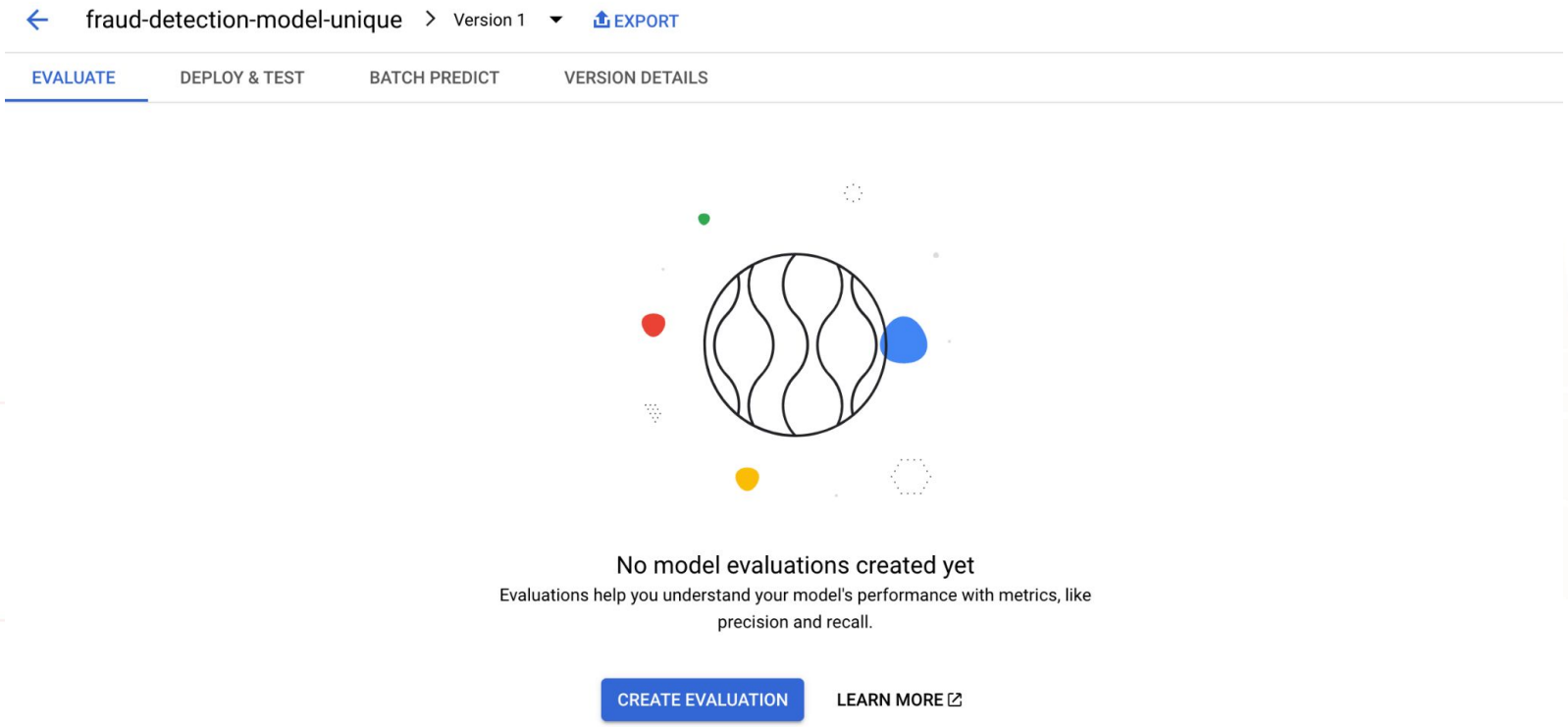
- Model Registry
- Online prediction
- Batch predictions
- Vector Search

Source: XXX

What is Model Registry?

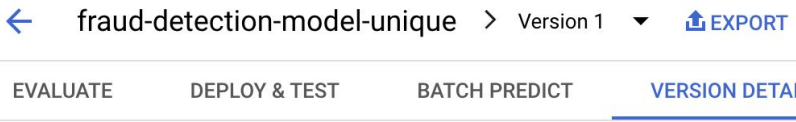


What is Model Registry?

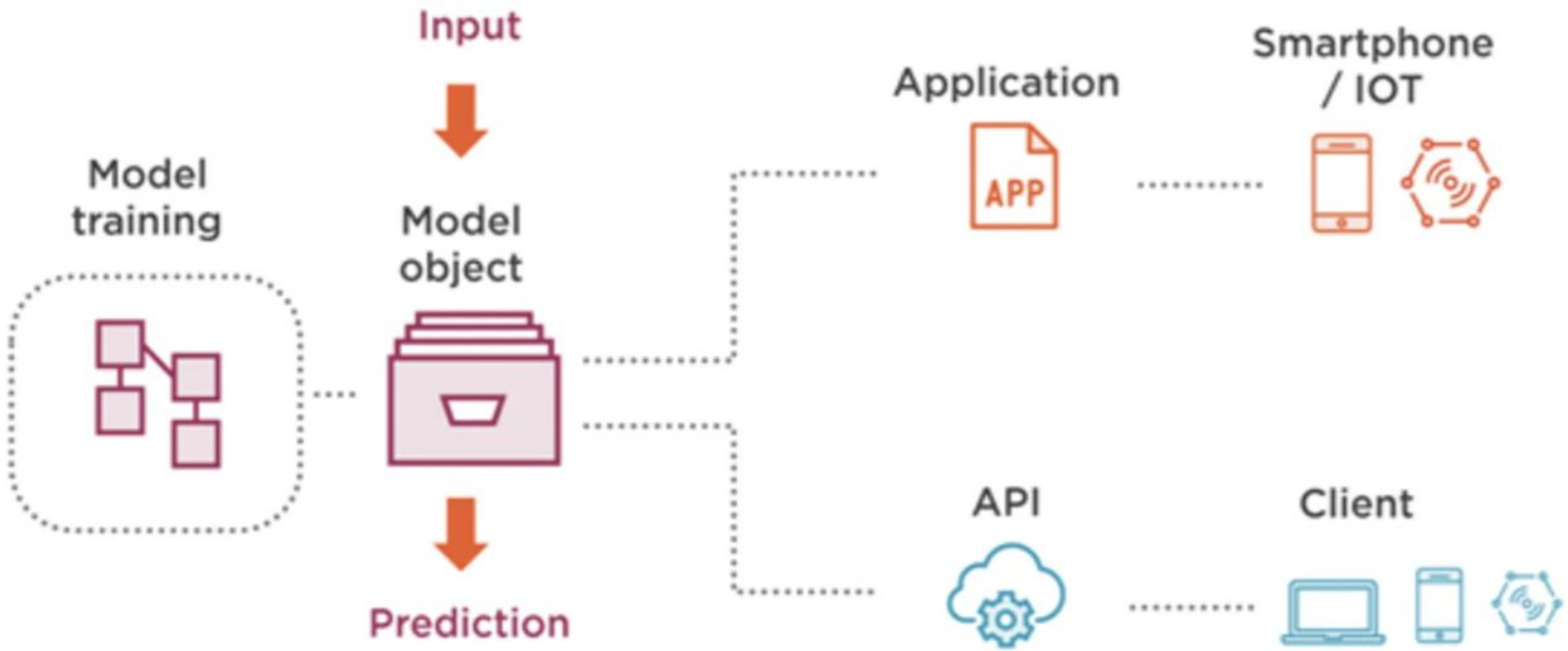


Source: XXX

Create an endpoint

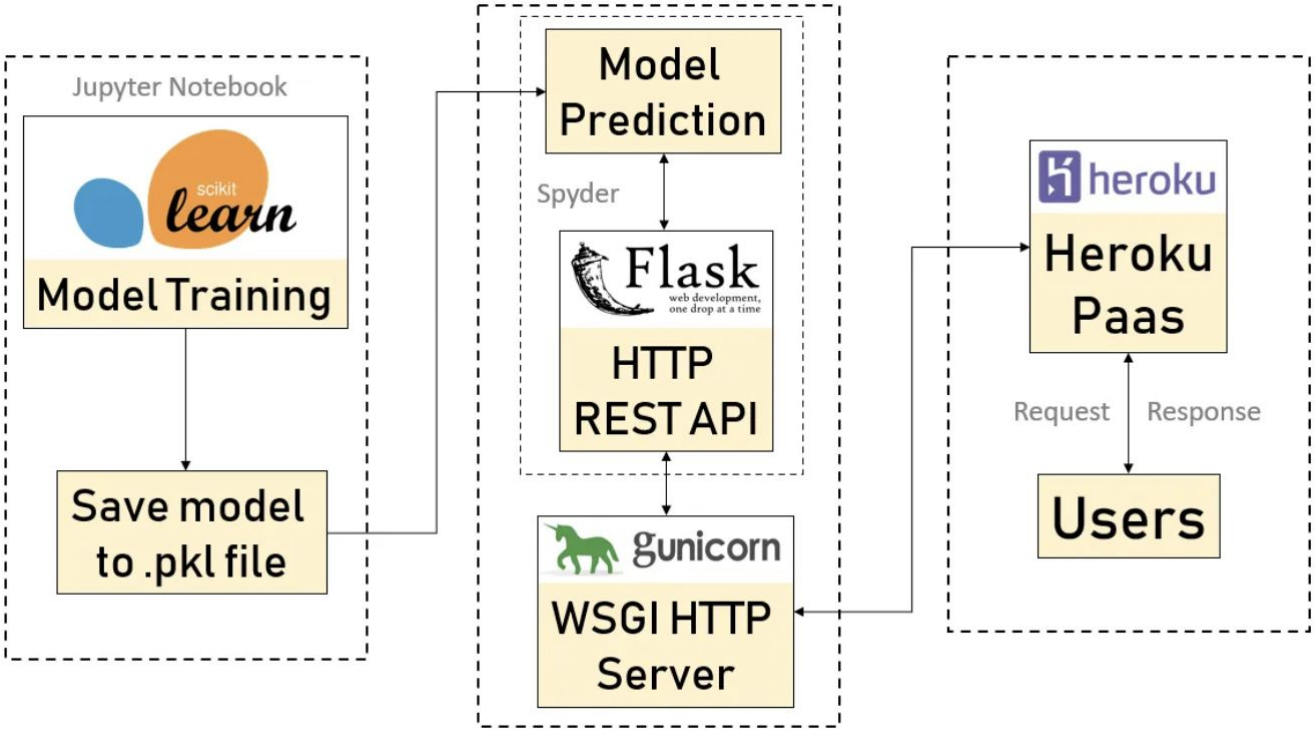


How do users interact with my model?



Source: https://ubuntu.com/blog/guide-to-ml-model-serving

What is an endpoint?



Source: https://medium.com/emilyshasha/machine-learning-model-deploy-on-website-using-flask-furniture-price-prediction-39a0d5ad41

Endpoints

```
# set the endpoint display name
ENDPOINT_DISPLAY_NAME = "fraud-detect-endpoint-unique"  # @param {type:"string"}
# create the Endpoint
endpoint = aiplatform.Endpoint.create(display_name=ENDPOINT_DISPLAY_NAME)
# print the endpoint display name
print("Display name:\n", endpoint.display_name)
# print the endpoint resource name
print("Resource name:\n", endpoint.resource_name)
```

```
Creating Endpoint
Create Endpoint backing LRO: projects/961778279858/locations/us-central1/endpoints/8169584699631468544/operations/3881366
687039094784
Endpoint created. Resource name: projects/961778279858/locations/us-central1/endpoints/8169584699631468544
To use this Endpoint in another session:
endpoint = aiplatform.Endpoint('projects/961778279858/locations/us-central1/endpoints/8169584699631468544')
Display name:
 fraud-detect-endpoint-unique
Resource name:
 projects/961778279858/locations/us-central1/endpoints/8169584699631468544
```

Endpoint

Deploy model to the endpoint

Set the following parameters for endpoint deployment:

- `endpoint` : The Vertex AI Endpoint resource created in the last step.
- `deployed_model_display_name` : Display name for the model. If not provided, model's display name is used.
- `machine_type` : Machine type required for serving the model on the endpoint.

```
# set the display name for the deployed model
DEPLOYED_MODEL_NAME = "fraud-detection-deployed-model"
# set the machine type for the endpoint
MACHINE_TYPE = "n1-standard-2"
```

Testing the endpoint

Source: https://katalon.com/resources-center/blog/top-5-free-api-testing-tools

Delete resources

- Undeploy the model
- Delete endpoint
- Delete model from Model Registry
- Delete from Online Predictions
- Delete bucket in Cloud Storage
- Delete Compute Engine instance
- Delete Workbench instance (notebook)

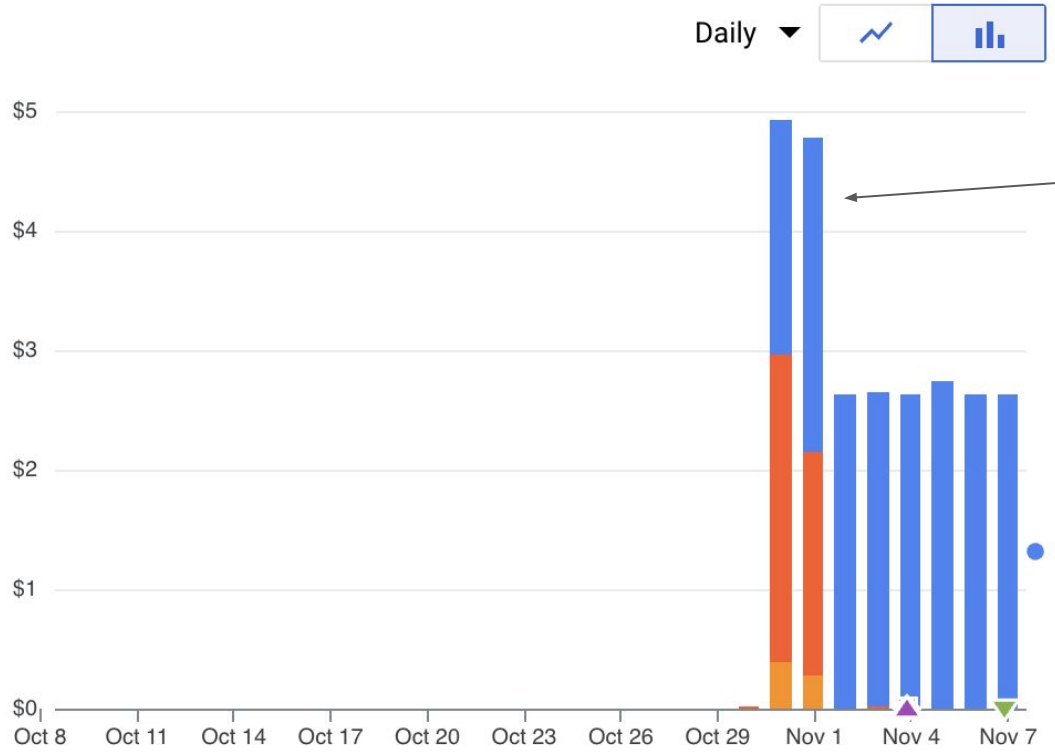
Learning Objectives for Today Review

1. Learn about Cloud History and its practical use cases for Machine Learning and AI applications

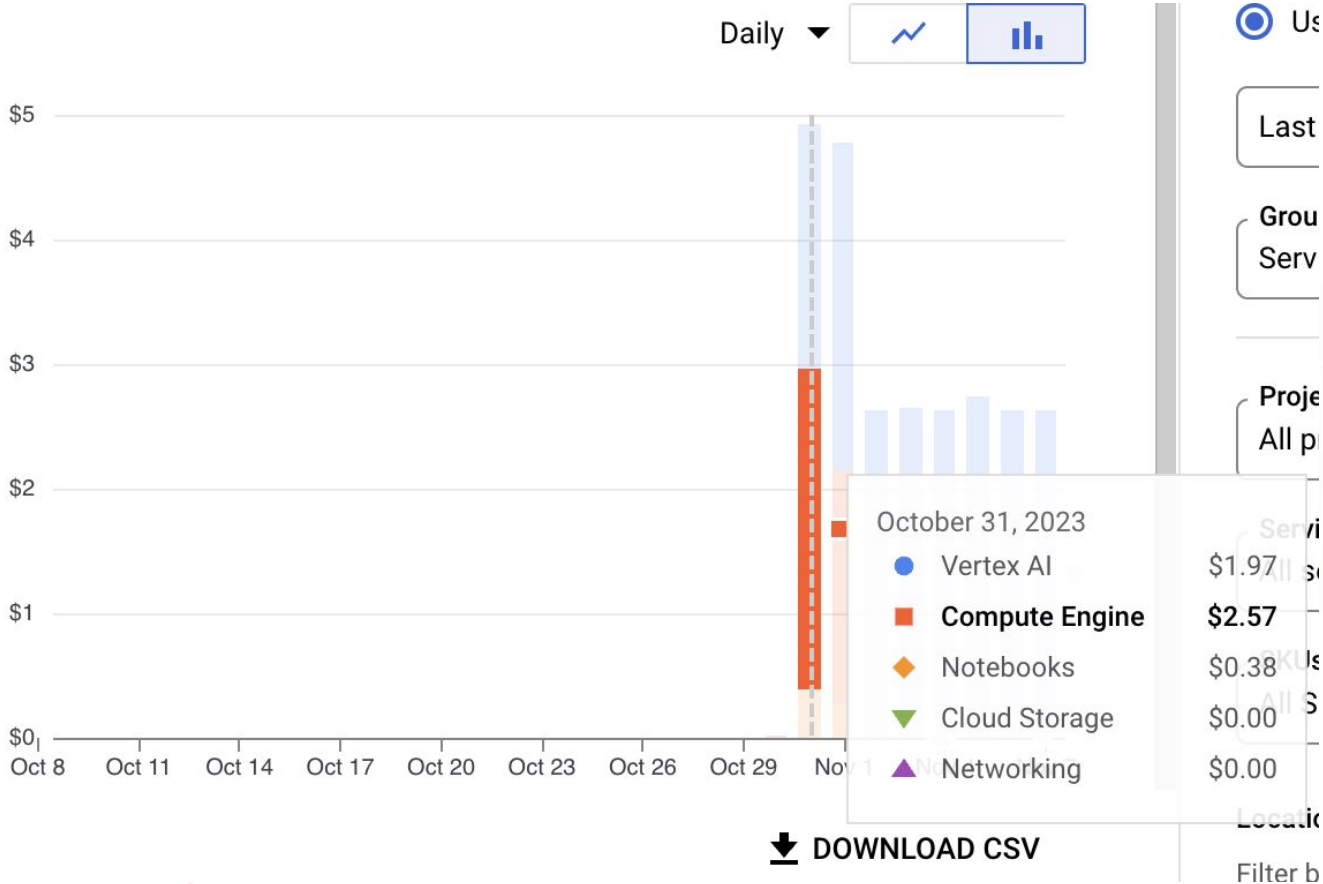
2. Understand the Machine Learning Lifecycle as it pertains to google cloud platform

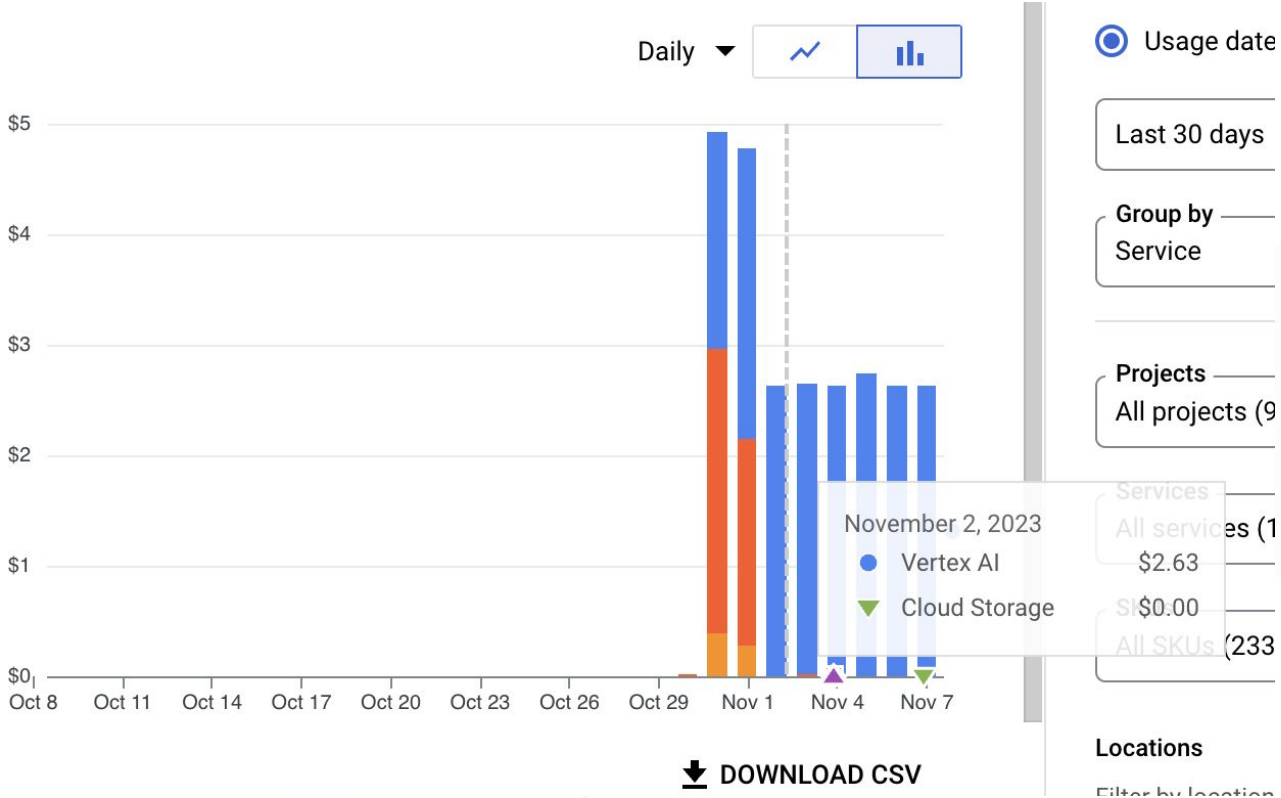
3. Apply understanding in a live demo detecting fraud for a large financial institution using Machine Learning

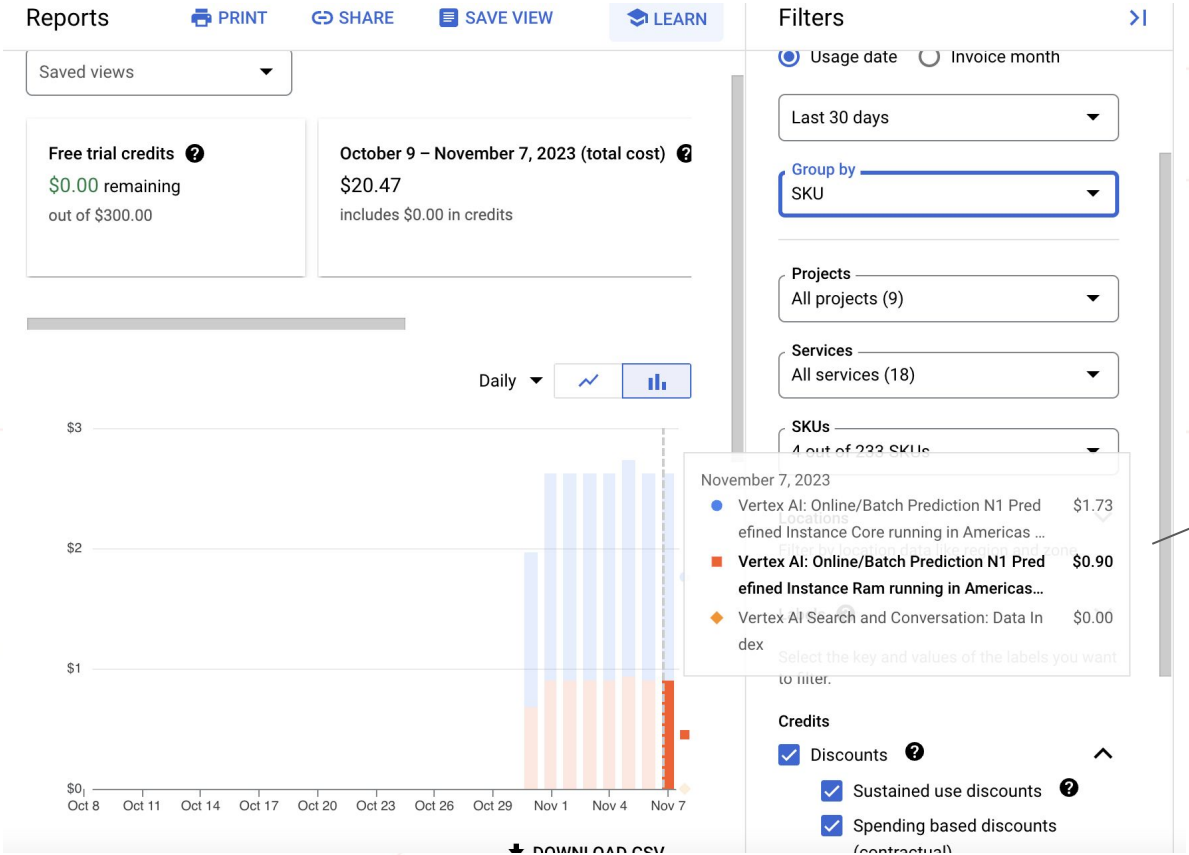
Further Questions?



Shut down notebook in Workbench, still had resources running in background


Daily
$5
$4
$3
$2
$1
$0
Oct 8
Oct 11
Oct 14
Oct 17
Oct 20
Oct 23
Oct 26
Oct 29
October 31, 2023
Vertex AI $1.97
Compute Engine $2.57
Notebooks $0.38
Cloud Storage $0.00
Networking $0.00
DOWNLOAD CSV
Last
Grou
Serv
Proje
All p
Filter b

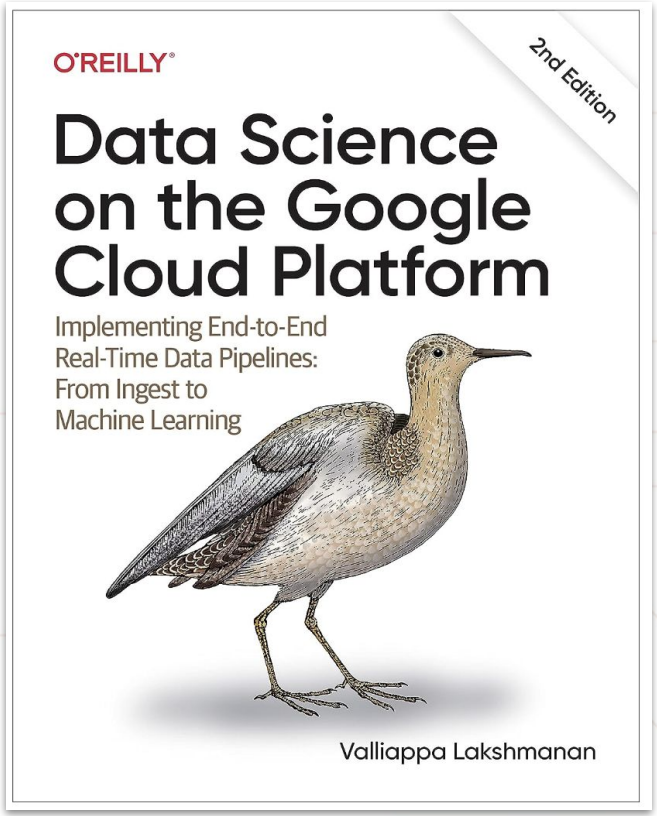

Daily
$5
$4
$3
$2
$1
$0
Oct 8
Oct 11
Oct 14
Oct 17
Oct 20
Oct 23
Oct 26
Oct 29
Nov 1
Nov 4
Nov 7
November 2, 2023
Vertex AI $2.63
Cloud Storage $0.00
DOWNLOAD CSV
Usage date
Last 30 days
Group by
Service
Projects
All projects (9
Services
All services (1
All SKUs (233
Locations



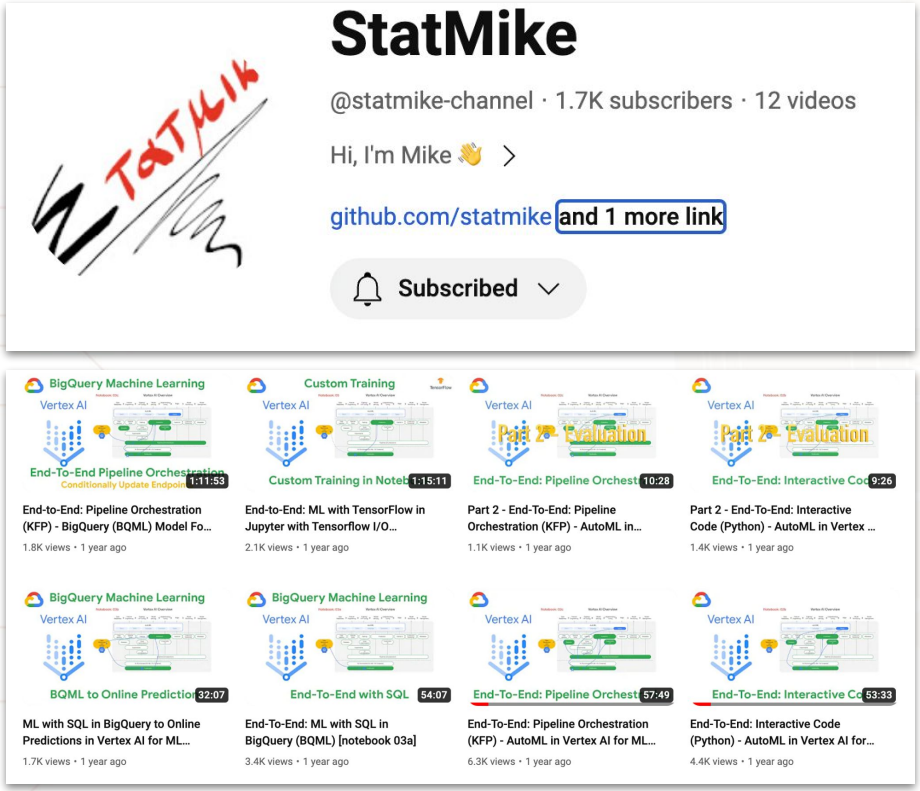
SKU

- Vertex AI: Online/Batch Prediction N1 Predefined Instance Core running in Americas for AI Platform
- Vertex AI: Online/Batch Prediction N1 Predefined Instance Ram running in Americas for AI Platform

Resources to continue your journey



Amazon Link



YT Link

Level up!


Google Cloud Certified
PROFESSIONAL
MACHINE LEARNING ENGINEER

Next Steps: Pipelines

Intro

https://medium.com/google-cloud/google-vertex-ai-the-easiest-way-to-run-ml-pipelines-3a41c5ed153

More advanced

https://blog.ml6.eu/a-general-framework-for-machine-learning-pipelines-on-gcp-b57e234f7d12

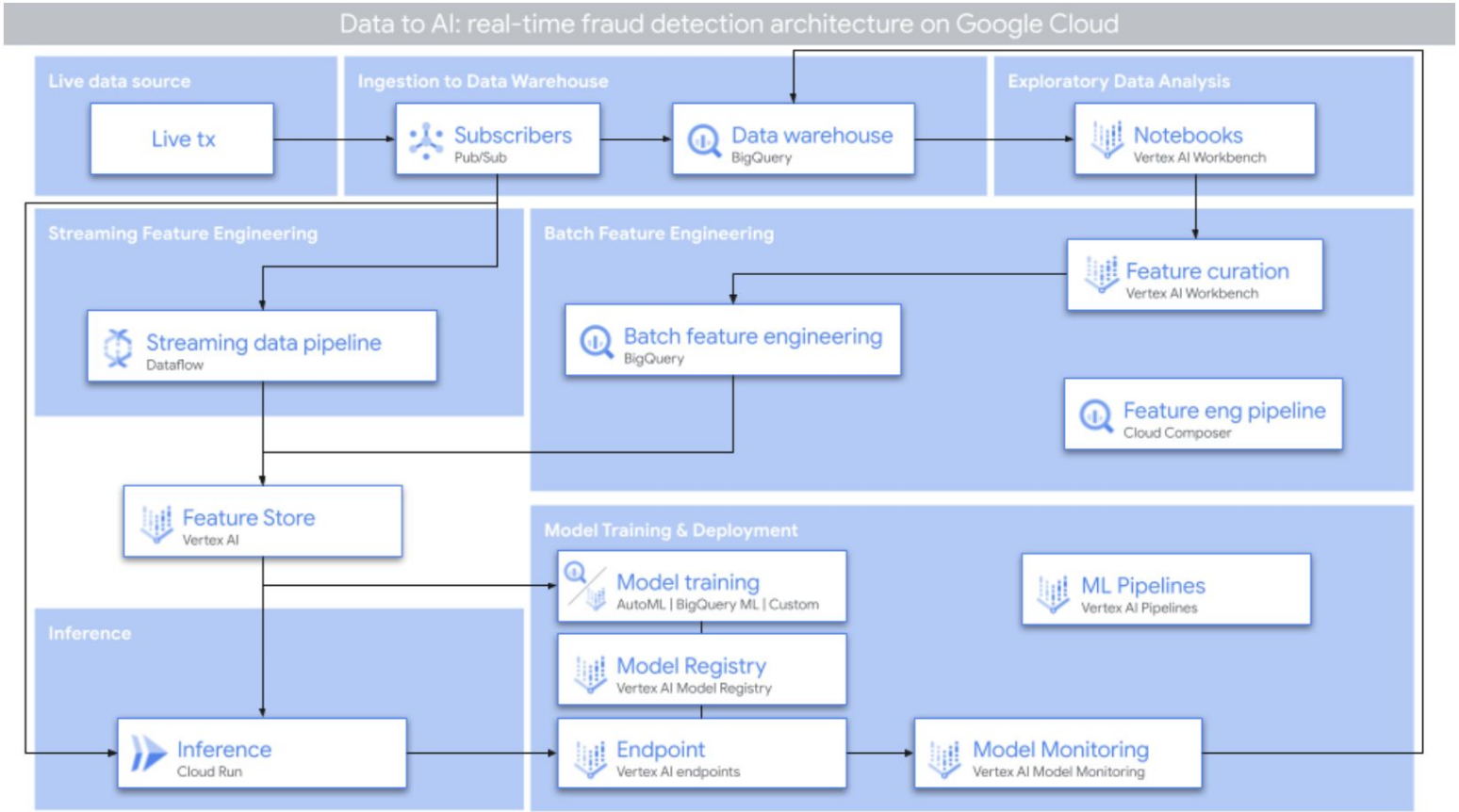
Google Vertex AI supports two different types of pipelines:

- Kubeflow Pipelines using the Kubeflow SDK ≥ 1.6
- TensorFlow Extended Pipelines using the TFX SDK ≥ 0.30.0

Pipelining Basics

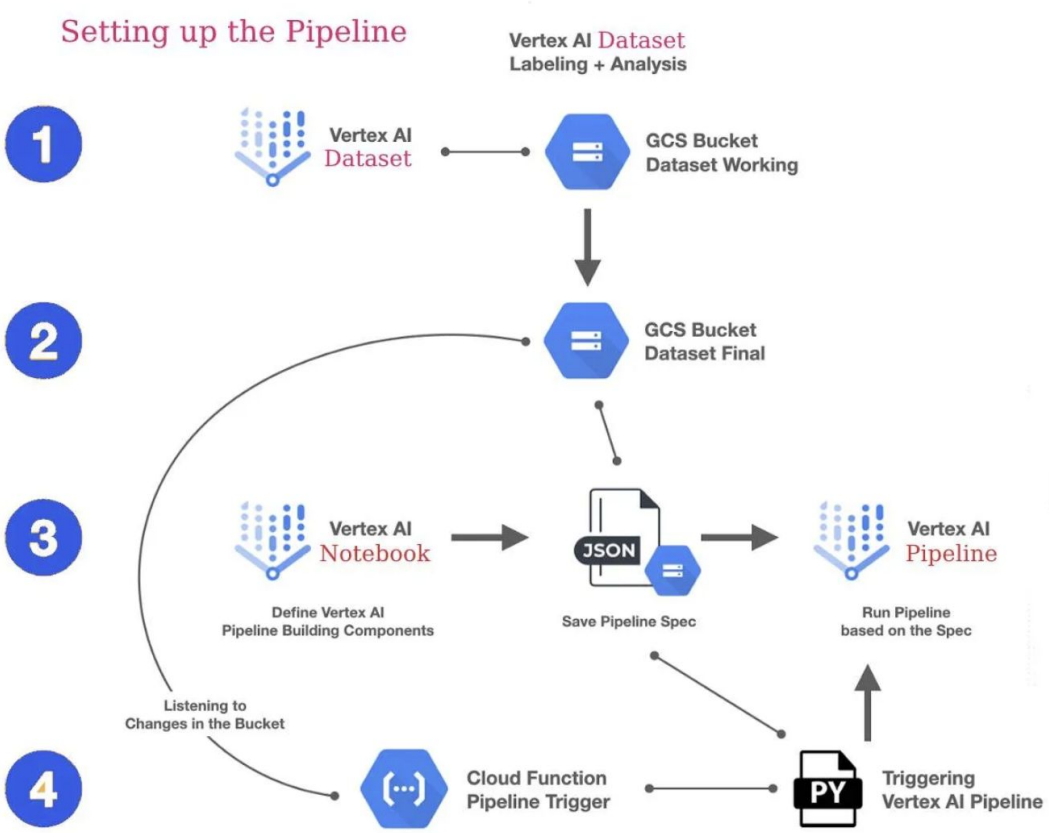
- Each component has its own compute resources
- A data preprocessing component doesn't need to have a GPU, but a deep learning model training component does
- Can customize to create base images. Think if you need to run a specific algorithm and utilize all cores. You wouldn't need this for the preprocessing step, but for XGBoost you would want to set jobs=-1
- Can monitor resource consumption on the component level

Next Steps: MLOps



Source:
https://www.googlecloudcommunity.com/gc/Architecture-Framework-Community/Build-an-end-to-end-data-to-AI-solution-on-Google-Cloud-with/ba-p/595682

Example MLOps Workflow



Source: https://blog.searce.com/implementing-mlops-pipeline-in-vertex-ai-to-adapt-to-the-changes-in-data-4cc3c07d5661